HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Electronics, Communication and Automation
Department of Signal Processing and Acoustics

Mika Ristimäki

Distributed Microphone Array System for Two-Way Audio Communication

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, June 15, 2009

Supervisor: Professor Matti Karjalainen
Instructors: M.Sc. Matti Hämäläinen

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Mika Ristimäki
Name of the thesis:	Distributed Microphone Array System for Two-way Audio Communication
Date:	June 15, 2009 **Number of pages:** 66
Faculty:	Electronics, Communication and Automation
Professorship:	S-89
Supervisor:	Prof. Matti Karjalainen
Instructors:	M.Sc. Matti Hämäläinen

In this work a distributed microphone array system for two-way audio communication is presented. The goal of the system is to locate the dominant speaker and capture the speech signal with highest possible quality. In the presented system each microphone array works as a Polynomial Beamformer (PBF) thus enabling continuous beam steering. The output power of each PBF beam is used to determine the direction of the dominant speech source. Finally, a Spatial Likelihood Function (SLF) is formed by combining the output beam powers of each microphone array and the speaker is determined to be in the point that has highest value of SLF. The audio signal capture is done by steering the closest microphone array to the direction of the speaker.

The presented audio capture front-end was evaluated with simulated and measured data. The evaluation shows that the implemented system gives approximately 40 cm localization accuracy and 15 dB attenuation of interference sources. Finally the system was implemented to run in real-time in the Pure Data signal processing environment.

Keywords: Microphone arrays, Localization, Pure Data, Audio Communication, Distributed signal processing

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Mika Ristimäki
Työn nimi:	Hajautettu mikrofoniryhmäjärjestelmä kahdensuuntaisessa äänikommunikaatiossa
Päivämäärä:	27.9.2009 **Sivuja:** 66
Tiedekunta:	Elektroniikka, tietoliikenne ja automaatio
Professuuri:	S-89
Työn valvoja:	Prof. Matti Karjalainen
Työn ohjaajat:	DI Matti Hämäläinen

Tässä työssä esitellään hajautettu mikrofoniryhmäjärjestelmä kahdensuuntaisessa äänikommunikaatiossa. Järjestelmän tavoitteena on paikallistaa hallitseva puhuja ja tallentaa puhesignaali mahdollisimman korkealaatuisesti. Työssä esiteltävässä järjestelmässä jokainen mikrofoniryhmä toimii polynomirakenteella parametrisoituna keilanmuodostajana (PBF), joka mahdollistaa jatkuvan keilanohjauksen. Hallitsevan puhelähteen suunta päätellään PBF:n jokaisen keilan ulostulotehoista. Lopuksi yhdistämällä jokaisen PBF:n kaikkien keilojen ulostulotehot muodostetaan avaruudellinen todennäköisyysfunktio (SLF), jonka suurin arvo määrää puhujan paikan. Puhesignaali tallennetaan ohjaamalla puhujaa lähinnä olevan PBF:n keila puhujan suuntaan.

Tässä työssä esiteltävän järjestelmän toiminta arvioitiin simuloidulla ja mitatulla datalla. Arvionti näyttää, että toteutettu järjestelmä pystyy paikallistamaan puhujan noin 40 cm paikannustarkkuudella ja järjestelmä vaimentaa muista suunnista tulevia häiriölähteitä noin 15 dB. Lopuksi järjestelmä toteutettiin reaaliakaisena systeeminä Pure Data signaalinkäsittelyympäristössä.

Avainsanat: Mikrofoniryhmät, Keilanmuodostus, Lokalisaatio, Pure Data, Äänikommunikaatio, Hajautettu signaalinkäsittely.

Acknowledgements

This Master's thesis has been done in Nokia Research Center (NRC) in Helsinki. First I would like to thank Martin Schrader and Jyri Huopaniemi for giving me a very valuable opportunity to work in one of the best industrial research centers. The experience has been unforgettable. My warmest gratitude goes for my instructor Matti Hämäläinen for giving me a deep insight into the world of array signal processing and helping with all the basics that I've supposedly have learned while drowsing in the university lectures. Also, I would like to thank my supervisor Professor Matti Karjalainen for several valuable advices.

I wish to extend my arm in appreciation for all my colleagues at NRC especially for Julia Turku for sharing the office with me for over an year and giving me tips for my thesis and helping with many other things, Johan Kildal for providing the atmosphere a little bit of South-European flavor, Riitta Väänänen for helping me with a lot of practicalities in my first days of work and last but not least I would like to thank Jarmo Hiipakka for giving me insight into some Matlab magic. I also would like to thank the Pasi Pertilä, Teemu Korhonen and Antti Löytynoja from Tampere University of Technology Audio Research Group for helping me with the measurements.

Finally I would like to thank my family, especially my Dad Hemmo and Mom Ulla who have been extremely supportive and caring in everything I have chosen to do in my whole life.

The last, but the most valuable thank you from all my heart goes to my beautiful and intelligent wife Isabel, who has been taking care of with love all the late hours that I have worked with my thesis. Te amo de todo mi corazón...

Espoo, June 15, 2009

Mika Ristimäki

Contents

Abbreviations

AEC	Acoustic Echo Control
RES	Residual Echo Suppression
DMAS	Distributed Microphone Array System
DOA	Direction Of Arrival
DSB	Delay-and-Sum Beamformer
FIR	Finite Impulse Response
GCC	Generalized Cross-Correlation
GEM	Graphics Environment for Multimedia
GSC	Generalized Sidelobe Canceller
HRTF	Head Related Transfer Function
LCMV	Linearly Constrained Minimum Variance
MMSE	Minimum Mean Square Error
MSE	Mean Square Error
PBF	Polynomial Beamformer
Pd	Pure Data
QMF	Quadrature Mirror Filter
RMSE	Root Mean Square Error
RTCP	Real-Time Control Protocol
RTP	Real-time Transfer Protocol
SACS	Spatial Audio Communication System
SLF	Spatial Likelihood Function
SNR	Signal-to-Noise Ratio
STD	Standard Deviation
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VAD	Voice Activity Detector
VBAP	Vector Base Amplitude Panning
WFS	Wave Field Synthesis

Chapter 1

Introduction

One of the key elements in our evolution to humans has been the ability to convey information by speech. It has been the root of literacy and thus it has given us the ability to pass our knowledge forward in easy and energy-efficient manner.

In today's globalizing world communication over large distances is increasingly important. Also due to the increasing awareness of the environment and global warming it is desired to use telecommunication methods instead of traveling in order to convey information. Although during the internet era new communication methods have emerged, such as e-mail and instant messaging, in most cases speech, or more generally, audio communication can still be considered the preferred way of communication. However, most of the present-day audio communication systems are still unable to capture and reproduce the audio signal in such way that the communication process would seem as natural as face-to-face conversation. Usually these single-channel audio communication systems do not use the full frequency range of human auditory system in audio capture or playback, the signal is corrupted with noise and reverberation and normal conversation flow is disturbed by long end-to-end delay of the system. In worst case scenario these problems make the received speech unintelligible. These problems are even more emphasized in a teleconferencing situation where many participants are present in the same environment.

To improve the audio communication quality, first the sound source should be captured so that the interferences are minimized in the captured signal. In teleconferencing situation the degradation happens mainly because of noise sources and reverberation inside the room. However, these effects can be reduced by using an array of microphones instead of just one microphone. A microphone array enables the use of *beamforming* [13, 75], where the sound source can be captured from specific spatial direction while at the same time the signals captured from other directions are attenuated. For that reason beamforming is often also called *spatial filtering*. More generally a microphone array processing can be seen

as spatial sampling, where an acoustical environment is not sampled only in time but also in various points in space. This provides more information about the spatial and temporal properties of the acoustical pressure wave.

The roots of microphone array signal processing are in antenna array theory [24, 33], where signal processing of multiple sensors was already a widely researched topic before the first microphone array applications by *Flangan et al.* over 20 years ago [29]. Since then microphone arrays have been applied for example in sound source localization [21], speech enhancement [22], speech recognition [87], hearing aids [67], blind source separation [14], surround sound recording [53], etc.

Microphone arrays can be designed by using several different design criteria depending on application, location, physical constraints, etc. The design can vary from huge microphone arrays of hundreds [71] or even over a thousand [85] microphones to small arrays of just a few closely spaced microphones. These small arrays are also the interest of this thesis. Also, various different geometric shapes have been used in the literature. The first type of microphone arrays were equally-spaced linear or rectangular arrays described in [29], where a microphone array is used to enhance the capture in large rooms. Also non-uniformly spaced and logarithmically spaced linear arrays have been researched for example in [70] and [78], respectively. In [65] Rafaely uses a spherical microphone array and spherical harmonics for plane-wave decomposition and in [47] a hemispherical microphone array is used.

To further improve parameter estimation and sound source capture, multiple microphone arrays can be placed at various different spatial locations inside a room, thus creating a Distributed Microphone Array System (DMAS). Because the intensity of a pressure wave is inversely proportional to the squared distance, the Signal-to-Noise Ratio (SNR) decreases the further a sound source is from a microphone array. By using a distributed system the effects of decreased SNR can be reduced, thus also improving the accuracy of parameter estimation.

In this thesis a distributed microphone array front-end to improve the spatial capture of two-way audio communication is presented. The goal of the system is to track the dominant sound source and to steer the beam of the nearest microphone array to the direction of the active speaker. The localization is done by measuring the energy of several possible speech source directions. The directions are defined by the outputs of Polynomial Beamformer [44]. Finally, the speaker location is estimated by fusing the sound source Direction Of Arrival (DOA) data of each individual microphone array. The audio capture can also be done by combining the beams from all the microphone arrays [41]. Other properties of a distributed system, such as time synchronization of each array and the array localization, are presumed known.

This thesis is organized so that in Chapter 2 a framework for a multi-channel audio communication system is presented and also the implementation of an experimental two-way audio communication platform is described. In Chapter 3 an overview of beamforming is given and relevant beamforming techniques are discussed in detail. Also, an algorithm for estimation of the speech source direction of arrival is presented and finally the fusing of the DOA data in distributed microphone array system is discussed. In Chapter 4 the measurements that were done to experiment with the proposed system are described and also the results of the experiments with simulated and measured data are shown. The conclusions are drawn in Section 5, where also the suggestions for future work are given.

Chapter 2

Spatial Audio Communication System

In this chapter the essential building blocks for a Spatial Audio Communication System (SACS) are discussed. First, in Chapter 2.1 an introduction to the SACS paradigm is given. In Chapter 2.2 audio capture methods for SACS are discussed and in Chapter 2.3 different methods for acoustic echo cancellation for SACS are described. Chapters 2.4 and 2.5 discuss the transmission and rendering of the audio data, respectively. A short overview for previous SACSs is given in Chapter 2.6 and finally an experimental SACS implementation is described in Chapter 2.7.

2.1 Introduction

To improve the quality and the naturalness of an audio communication system a multi-channel or Spatial Audio Communication System (SACS) is needed. A SACS consists of a multi-channel capture front-end to capture the acoustical wave field, audio compression to transmit the signal efficiently over a network and a multi-channel reproduction system to render the captured wave field. Without any additional notes, in this work SACS always refers to a two-way audio communication setup, where both ends of the communication chain can work as a transmitting and receiving end. A basic block diagram of a SACS is shown in Figure 2.1 as presented in [66].

2.2 Audio Capture

The purpose of the spatial audio capture or front-end of SACS is to capture the source signals without any deterioration in the sound quality due to multi-path propagation or external

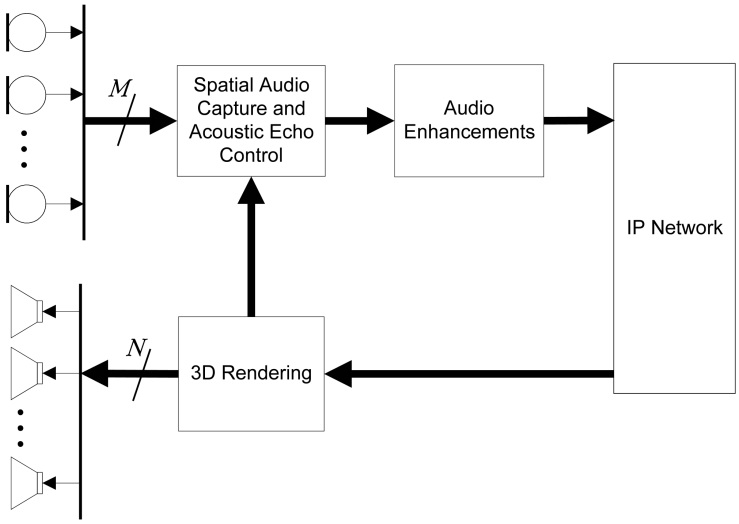


Figure 2.1: A Spatial Audio Communication System (SACS)

noise sources. Also the capture system is used to estimate the sound source locations and to capture the surrounding ambience signal. An ideal way to capture the source signals would be the use of blind signal separation algorithms [57]. In blind signal separation multiple mixed source signals are separated from each other, without *a priori* information about the signals. However, instead of simple additive mixing, in acoustic signals the mixing process is convolutive and time-varying, which makes the blind signal separation methods not yet robust enough to work reliably under real-world time-varying conditions [23].

Therefore, the most common way to enhance the quality of the source signal capture is to use beamforming techniques. In beamforming the directivity pattern of the microphones is electronically steered so that the sensitivity of the microphones is largest to the direction of the desired signal sources. Additionally at the same time, beamforming enables the suppression of interfering sources under the condition that they are not at the same direction with the desired signals. More detailed description about beamforming is given in Chapter 3.1. A good overview of adaptive beamforming can be found also from [36].

2.3 Echo Cancellation

The biggest problem of SACS has been the lack of efficient solutions for the fundamental echo problem. When the far-end signal is played from N loudspeakers, all the M microphones capture the direct signals from the loudspeakers and also the reflections of theses signals from room structures. This undesired signal should be removed from the captured signal before it is sent back to the far-end of the SACS. This Acoustic Echo Cancellation

(AEC) problem is far from trivial to solve but essential for SACS performance. The basic principle of a single-channel echo canceller using adaptive filtering is shown in Figure 2.2. As can be seen from the figure, in AEC techniques the impulse response of the echo path between the loudspeaker and the microphone is estimated by using adaptive filtering and this estimation is used to filter the far-end signal. The filtered signal is subtracted from the captured signal ideally producing an echo-free signal. However, in multi-channel case there are $N \times M$ echo paths, which in itself makes the Multi-Channel AEC (MC AEC) computationally very expensive. Additionally, all the loudspeaker signals are usually from the same source and thus highly correlated. This so called non-uniqueness problem prevents or dramatically slows the convergence of adaptive estimation of the acoustic echo paths when the single-channel AEC algorithms are applied to a MC AEC case [73]. However, some solutions have been proposed for the adaptive MC AEC problem as presented e.g. in [8, 16, 73]. Because the non-uniqueness problem happens due to the correlated loudspeaker signals, one of the most used methods in MC AEC is to try to decorrelate the loudspeakers signals. This decorrelation can be done e.g. by adding independent noise signals to the loudspeaker signals, using decorrelation filters or interleaving comb filters [73]. Also, MC AEC techniques in conjunction with beamforming have been presented e.g. in [15, 40].

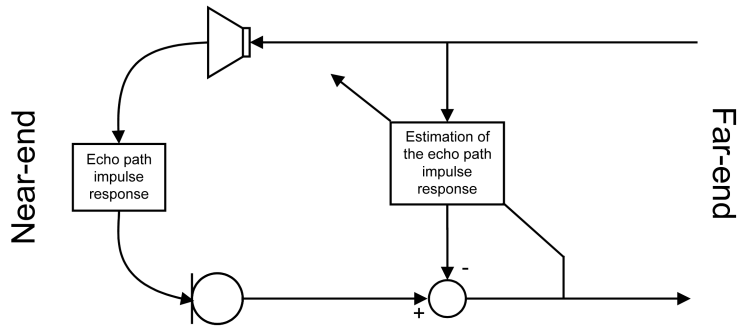


Figure 2.2: A single-channel adaptive echo canceller

When using high sampling rates for high quality SACS, another problem with echo cancellation rises that is due to the length of the required echo cancellation filter. The filter length can increase so that efficient filtering becomes computationally heavy and slowly converging. For example when sampling rate of 32 kHz is used in a moderately reverberant room ($T_{60} = 0.5s$), the filter should be 16000 taps long to attenuate all the echo. However, in [80] *Wallin et al.* suggest a hybrid AEC and Residual Echo Suppression (RES) to avoid the use of long AEC filters. They suggest that microphone signal is divided into two subbands and then adaptive filtering is used only for the lower subband and suppression techniques for the higher subband. The acoustical echo suppression techniques are compu-

tationally much lighter than adaptive filtering but on the other hand sometimes suppression introduces audible distortions. Therefore *Wallin et al.* suggest that it is enough to use suppression in the higher frequency band because most of the acoustic energy is in the lower subband. This way the computational complexity of the AEC system can be kept low while maintaining the perceived sound quality high. Also, it has to be noted that it is not always necessary to remove all the echo and for some applications, such as acoustic opening [32], only the removal of direct sound may be sufficient.

2.4 Audio Streaming

Nowadays the most popular way to transmit audio data is to do it over an Internet Protocol (IP) network. Two main protocols for audio data transfer over IP are Transmission Control Protocol (TCP) [59] and User Datagram Protocol [58] (UDP). The main difference between TCP and UDP is that TCP is connection-oriented protocol that guarantees that the sent data packet is received. On the other hand UDP is connectionless protocol that only sends the packet without concern if the packet is received or not. Because of received packet confirmations and resending, TCP increases the end-to-end latency of the communication system, which is undesired for real-time communication. Moreover, occasional missing audio packets do not reduce the communication system quality significantly, which makes UDP more suitable for real-time audio data transfer protocol. However, because of its unreliable nature, UDP in itself can be insufficient for robust audio data transfer, and therefore Real-Time Transfer Protocol (RTP) [68] and Real-Time Control Protocol (RTCP) can be used on top of UDP to provide more reliable end-to-end transmission, while maintaining low latency performance.

Although broadband internet connections are increasingly popular, a CD quality audio data can still be considered to have a relatively high network bandwidth consumption. To reduce the bit-rate of the audio signals and thus enable the usage of the system under low bandwidth conditions, an audio compression method is needed. However, it should be noted that the algorithmic latency of the compression method has to be low in order to maintain acceptable end-to-end latency in the communication system.

2.5 Audio Rendering

Similarly as in audio capture, the goal of spatial rendering is to reproduce the far-end sound field in the listening location. However it should be noted that for all SACSs complete 360° spatialization is not always the desired property but sufficient spatialization of the far-end sound sources could be enough. This way other features of SACS can be optimized

such as the need for external hardware and mobility. Many different techniques exist to spatialize a sound field. For a wavefield reconstruction, Wave Field Synthesis (WFS) [9] has been proposed. WFS is based on the Huygens' principle, which states that any wave front can be created by using a superposition of elementary spherical waves. In WFS the elementary spherical waves are created by multiple loudspeakers. In practice however, several artifacts emerge [74] and nearly perfect wavefront reconstruction would require a very large loudspeaker array where the loudspeakers are placed only a few centimeters from each other.

Various amplitude panning methods have also been suggested. In amplitude panning the gain of each loudspeaker is modified in order to create virtual sound sources. Most common amplitude panning methods are Ambisonics (and its higher-order variants) [49] and Vector Base Amplitude Panning (VBAP) [63]. In Ambisonics the loudspeakers are usually placed in symmetrical setups and the sound source is played from all the loudspeakers simultaneously. In VBAP multiple loudspeakers are divided into pairs (2D) or in triplets (3D) and only one pair or triplet is used at a time to render the virtual sound source between the loudspeakers. VBAP can be used with arbitrary loudspeaker setups and the localization accuracy can be increased by increasing the number of loudspeakers. A short overview of multi-channel reproduction methods can be found in [64].

Also binaural technology [54] can be considered. Binaural technology is based on a set of transfer functions, called Head Related Transfer Functions (HRTF), that are estimated from different source directions to the ear canal. HRTFs can then be used to create virtual sources around the listener. Binaural technology is mainly suitable for headphone use, where the sound is played directly to the ear canal, because the estimation of HRTFs is also made in the ear canals. However, binaural playback techniques for stereo loudspeakers also exist. An overview of HRTF techniques for 3D audio and virtual acoustics can be found e.g. in [39].

2.6 Spatial Audio Communication Systems

Although a successful commercial SACS is still yet to come, it has been under active research for two decades. One of the first teleconference systems to adopt the SACS principles was a two-channel stereophonic audio communication system presented by Botros *et al.* in [12]. Even though their system was not fully multi-channel system, they showed that a stereophonic communication system improved the speech intelligibility and made the localization of the far-end speaker possible. Another type of two-channel audio communication system utilizing binaural technology was presented in [86]. In this so called binaural telephony the microphones are placed in the ears of a listener or a dummy head and the cap-

tured sound field is also reproduced using headphones. In [25] Evans *et al.* compared the binaural reproduction system with multi-channel loudspeaker reproduction system using the Ambisonics technique. Evans *et al.* argue that SACS utilizing the Ambisonics method will always have a very narrow market, because of Ambisonics' extensive need for external hardware (loudspeakers and amplifiers), fixed placement of the loudspeakers and the systems need for calibration, thus significantly reducing the portability of the communication system. The first extensive SACS framework was proposed by Herbordt et al. [34] where most of the research areas, excluding multi-channel audio coding, concerning the SACS paradigm are discussed and also a SACS using adaptive beamforming, multi-channel echo cancellation and Wave Field Synthesis for rendering is presented. A similar approach is presented in [32] where also various capture methods are compared.

2.7 Implementation of the Experimental Audio Communication Platform

An audio communication system testing and validation is difficult to do with just using simulated conversational setups, and therefore a real-time two-way audio communication system is needed. In this chapter an implementation of a Spatial Audio Communication System (SACS) made on top of Pure Data (Pd) audio and video signal processing software is presented. The details of the experimental audio communication system are provided in the Chapters 2.7.1 and 2.7.2.

2.7.1 Pure Data Environment

Pure Data[1] is a cross-platform open-source software that is used for audio, video and graphical signal processing [61]. It was originally developed by Miller Puckette for audio signal processing. However, after its release also video and graphical signal processing extensions have been added to Pd.

Pure Data can be considered as a graphical programming environment. The starting point in Pd is an empty patch or canvas, into which the developer implements the signal processing algorithm. The algorithms are developed by using elemental signal processing tasks (e.g. add, multiply, cosine, oscillator) called objects, that are connected to other signal processing objects with patch cords. By creating networks of these signal processing objects, complex algorithms can be implemented. An example of and Pd patch can be seen in Figure 2.3 where a simple frequency modulation algorithm is presented.

New signal processing objects or so called *Pd externals* can also be written in C/C++

[1] http://puredata.info

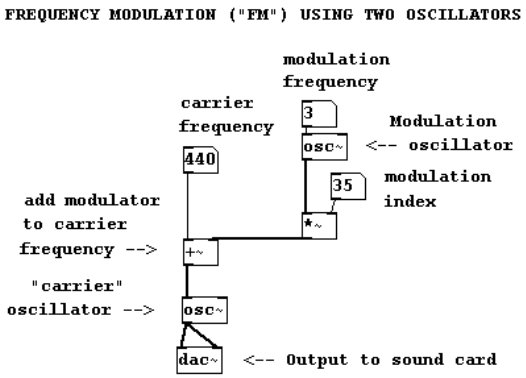


Figure 2.3: A basic Pd patch for frequency modulation

using well known software development practices that enable algorithm portability also to other software platforms. Pd has a well defined interface for extending its capabilities. As Pd is an open-source software, it has created an active developer and user base which supports the future maintenance and development of the environment.

Pure Data was chosen to be the backbone of the communication system because it can be easily interfaced with high-quality multichannel sound cards and other external hardware modules. Also, with Pd the developer does not have to deal with audio hardware drivers or I/O interrupts but he/she can concentrate strictly on the algorithm development.

However as Pd was originally developed for computer music purposes [62] and it was never intended to be a real-time algorithm development environment, it does not provide analytical algorithm development tools. Therefore MATLAB was integrated with Pd to analytically test and validate the real-time algorithm performance [66].

2.7.2 Two-way Communication System

Experimenting audio communication systems with just simulated conversational setups does not give real information about the performance of the system. Simulated setups usually lack the real dynamically changing environment and therefore all the underlaying problems may not be correctly understood and assessed. This also makes the creation of a feasible test material more cumbersome.

Therefore it was decided to implement a high-quality (sampling rate 32 kHz) real-time SACS for better evaluation of the SACS performance and user experience. To extend the capabilities of the system, various previously implemented DSP algorithms were ported to Pd, using Pd's C/C++ interface. A block diagram of the implemented SACS can be seen in Figure 2.4.

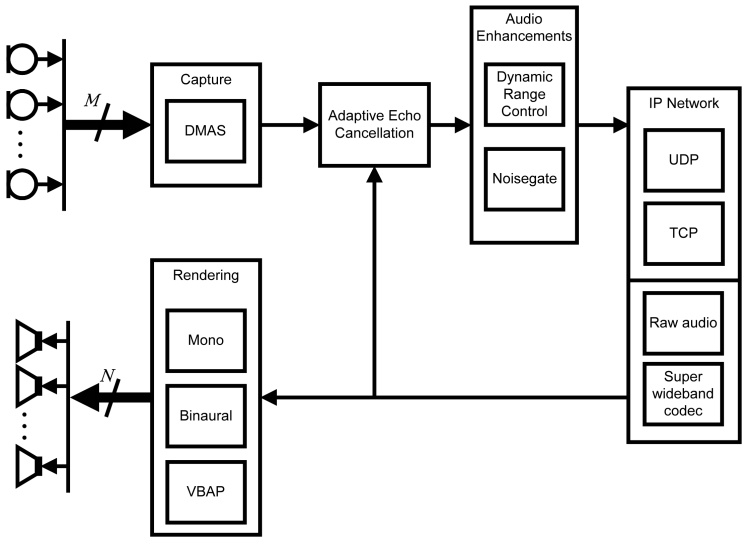


Figure 2.4: The block diagram of the implemented communication system

As can be seen from the block diagram, various capture and rendering methods have been implemented. The audio capture can be done in single-channel mono format or with the Distributed Microphone Array System (DMAS) presented in Chapter 3. Similarly, audio rendering can be done in single-channel format and two different spatial rendering methods have been implemented. Binaural rendering is intended for headphone use but can also be used for closely placed loudspeakers. For multiple loudspeakers, Vector Base Amplitude Panning (VBAP) can be used as a rendering method.

The communication is done by transmitting single-channel audio stream over IP network. The audio transfer is based on *netsend~/netreceive~* [3] Pd externals made by Olaf Matthes. These externals enable the multichannel transfer of raw audio data using TCP and UDP. While TCP causes more latency to the end-to-end connection than UDP, in experiments it has been noted that in this audio transfer implementation TCP protocol provides more robust audio data transfer in the current network environment. Also, to reduce the required network bandwidth the externals were modified to support audio transfers using an super-wideband audio codec. The audio codec can be configured for low bandwidth situation by decreasing the data bitrate at the expense of audio quality. A video feed transmission was also implemented using Graphics Environment for Multimedia (GEM) extension for Pd and H.264 video codec.

2.7.3 Wide-band Acoustic Echo Control

As stated in Chapter 2, echo cancellation is essential for good SACS performance. For this SACS implementation a single-channel adaptive echo cancellation algorithm described in [55] was ported to Pd as a Pd external. However, because the sampling rate of the SACS was set to 32 kHz, an extremely long adaptive filter would have been required for sufficient echo attenuation. Therefore, a hybrid echo canceller was implemented as suggested in Chapter 2.3.

As shown in Figure 2.5 the hybrid echo canceller/suppressor was implemented by first dividing the input signal to two sub-bands of 0 - 8 kHz and 8 - 16 kHz. To further reduce the length of the required adaptive filter, the lower sub-band was still divided to two sub-bands of 0 - 4 kHz and 4 - 8 kHz. The adaptive filtering was used only for the lowest sub-band and suppression was used for the higher sub-bands. Previously implemented Quadrature Mirror Filter (QMF) [76] was ported to Pd and used to filter the signal to sub-bands.

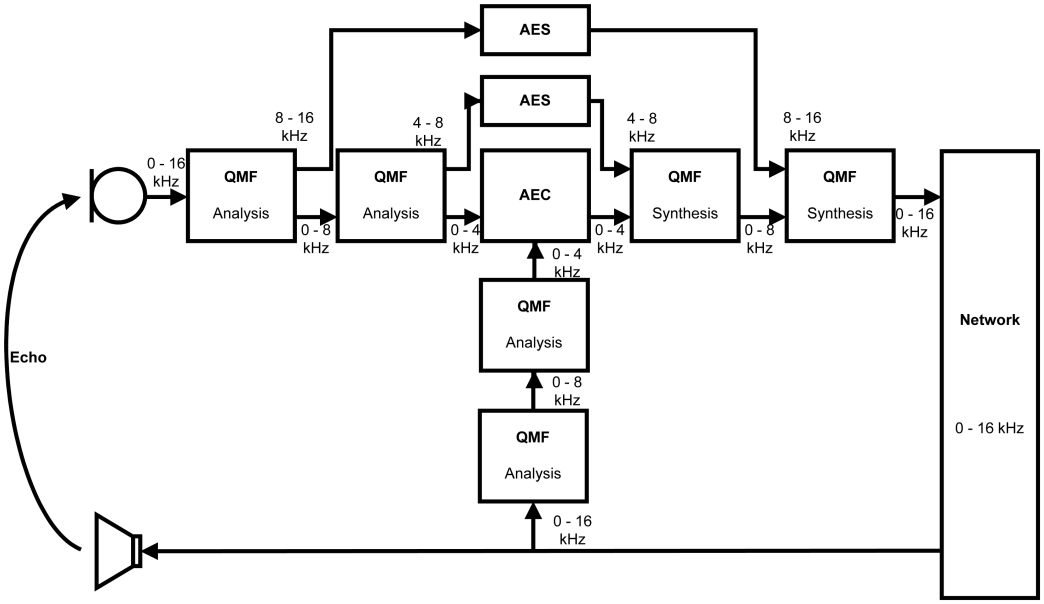
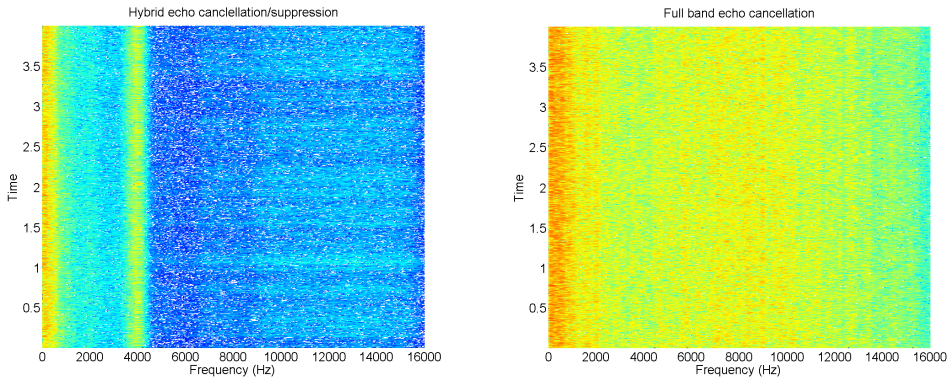


Figure 2.5: The hybrid adaptive echo cancellation/suppression

However, the use of QMF to filter the signals is not optimal for hybrid echo cancellation/suppression [30]. In QMF the bandpass filters overlap in the transition band in order to have minimal attenuation in all the frequency band. This overlap also extends the cutoff frequency of the bandpass filters and therefore the sub-band signals are not strictly bandlimited to half sampling rate. During the downsampling this causes aliasing in the transition band. This property of QMFs deteriorates the AEC performance as can be seen from Figure 2.6,

where in Figure 2.6(a) the echo is not attenuated from the transition bands. Figure 2.6(b) shows a spectrogram of a full-band AEC with the same filter length as in left image. In full-band AEC, filtering is used for the whole bandwidth.



(a) Hybrid AEC and RES

(b) Full band AEC

Figure 2.6: Spectrogram of the outputs of hybrid AEC and full band AEC. Input signal is white Gaussian noise.

The echo cancellation was placed after the DMAS for computational purposes. In [40] Hämäläinen and Myllylä compare different echo cancellation and beamformer integration schemas. In their paper they conclude that when echo cancellation is placed after the beamformer (so called 'AEC-last' configuration), the AEC performance deteriorates rapidly when even small changes occur in the steering direction of the beamformer. Therefore they suggest an 'AEC-middle' configuration, where the AEC is placed between the PBF pre-filter and the post-filter. However, due to the computational limitations this can not be done in real-time in the implemented SACS. Similar AEC and microphone array integration strategies are also proposed by Kellermann in [46].

However, the main objective of this work is presented in the following chapters, where the theory and evaluation of a Distributed Microphone Array System (DMAS) is described. The goal of the proposed DMAS design is to function as an audio capture front-end in the experimental audio communication platform described in this chapter.

Chapter 3

Microphone Array Techniques

In this Chapter the microphone array techniques for a Spatial Audio Communication System (SACS) front-end are described. In Chapter 3.1 theory of beamforming is reviewed and more detailed description of Polynomial Beamformer (PBF) is given. After that in Chapter 3.2 a framework for speech source Direction Of Arrival (DOA) estimation based on beamforming is presented. Finally in the last Chapter 3.3, the data fusion from several microphone arrays is discussed and a method for distributed speech source localization is presented.

3.1 Beamforming

In beamforming the directivity of a sensor array is altered in order to maximize the array sensitivity to the direction of the desired source while at the same time minimizing the array sensitivity in the direction of interfering noise sources. It can also be considered as spatial filtering where signals are separated by their physical location.

In the following, first an introduction to beamforming techniques is given in 3.1.1. Next, a review of the basic beamforming theory is given in Chapter 3.1.2 and measures for beamformer evaluation are presented in Chapter 3.1.3. Finally, the theory of Polynomial Beamformer (PBF) structure for continuous beam steering is described in Chapter 3.1.4.

3.1.1 Background

In general, beamforming techniques are widely used in antenna, sonar and radar applications. In principle, the idea in beamforming is to place FIR filters in each sensor channel thus enabling spatial and temporal control over the beamformer response. Beamforming techniques can be divided between *data independent* [83] and *data dependent* [79] beamforming. In data independent beamforming the signal processing is not dependent on the

statistics of the sensor data whereas in data dependent beamforming the estimated statistics of the sensor data it taken into account when optimizing the array response.

The first applications of the data independent beamforming techniques to wideband speech signals were made by Flanagan *et al.* [29], where they introduced a beamforming technique and microphone array design criteria for large auditoriums. However, their design was only effective for narrow band signals and therefore various generalizations to wideband signals have been proposed e.g in [19, 28, 81].

Most data dependent beamformer designs are based on so called Minimum Mean Square Error (MMSE) design or Linearly Constrained Minimum Variance (LCMV) design. MMSE beamformer design is based on minimizing the Mean Square Error (MSE) between desired response and the actual response of the beamformer. The main problem with MMSE design is that it requires estimates separately for the interference signal and the source signal cross-correlations, which in general can be difficult to estimate. Also, MMSE design does not guarantee undistorted output signal [36, 72].

In LCMV design the estimation of the desired signal is avoided by using the estimate of the overall microphone signal cross-correlations and imposing constraints on the estimated position of the desired source signal. The beamformer response can be derived by minimizing the beamformer output variance subject to the constraints. The most widely used LCMV beamforming techniques are based on the so called Generalized Sidelobe Canceller (GSC) that was introduced by Griffiths and Jim [31]. Various extensions and improvements to GSC have been proposed e.g in [35, 37].

Although GSC based beamformers perform well if the direction of the desired source is known, they are not suitable for situations where several source directions have to be taken into account. This is because the constraints are dependent on *a priori* known source locations. Also, slow convergence of the adaptive filtering used in most of the GSC based beamformers makes the beamformer response unpredictable in dynamically changing environments. Furthermore, in this work the beamformer output is used for speaker localization, which makes the use of data dependent beamforming unpractical. Therefore here the interest is in data independent beamformers and especially in Polynomial Beamformer (PBF) which parameterizes the beamformer FIR filter coefficients and thus enables continuous beam steering.

3.1.2 Basic Beamforming Theory

Consider a point source S at location $\mathbf{p}_s = [p_{s,x}\, p_{s,y}\, p_{s,z}]$ radiating harmonic spherical pressure waves in a lossless (no turbulence or temperature changes) and noisy environment. An arbitrary microphone array of M ideal omnidirectional microphones measures the pressure wave at $\mathbf{p}_m = [p_{m,x}\ p_{m,y}\ p_{m,z}]$, where $m = [1,\ 2,\ \ldots\ , M]$. In this work it is assumed

that the Euclidean distance between $\mathbf{p}_s$ and $\mathbf{p}_m$ is long enough in order to approximate the pressure wave with a plane wave. The output of m^{th} microphone in discrete time domain is

$$x_m(n) = s_m(n) + \vartheta_m(n), \tag{3.1}$$

where $s_m(n)$ is the source signal as captured by the m^{th} microphone and $\vartheta_m(n)$ is the noise signal in the m^{th} microphone. The source signal and the noise signal are assumed to be zero mean and uncorrelated.

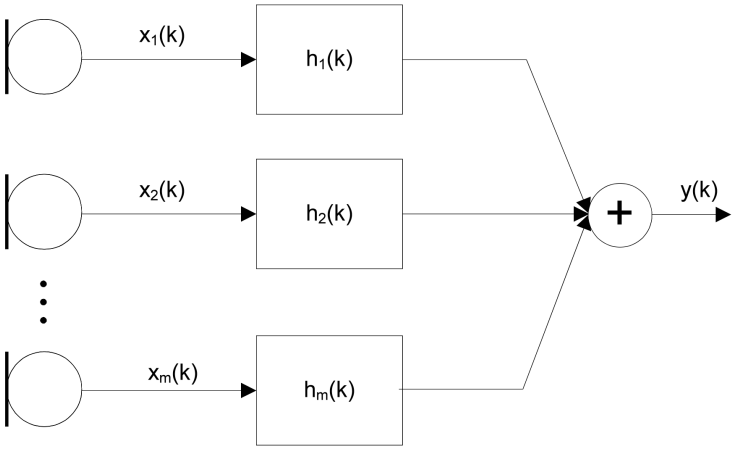


Figure 3.1: A basic beamformer structure

A basic *filter-and-sum* beamformer structure is shown in Figure 3.1. Thus, the output of the beamformer can be written as

$$y(n) = \sum_{m=1}^{M} \sum_{k=0}^{L-1} h_m(k) x_m(n-k), \tag{3.2}$$

where $h_m(k)$ is the impulse response of the m^{th} beamforming filter and L is the filter length. In the most trivial case the filters $h_m(k)$ can be just delays. In this case, the delays are adjusted so that the phase of the source signal is aligned to be the same at each microphone location. The output of this so called *Delay-and-Sum Beamformer* (DSB) [42] is

$$y(n) = \sum_{m=1}^{M} x_m(n - \tau_m), \tag{3.3}$$

where $\tau_m = (|\mathbf{p}_s - \mathbf{p}_m| - |\mathbf{p}_s - \mathbf{p}_1|)/c$.

The main problem with DSB is that the output beam pattern is frequency dependent. In high frequencies the beam is narrower than in the low frequencies. In low frequencies the wavelength is larger, thus resulting in a wider beam. This is demonstrated in Figure 3.2

where a response of a equally-spaced linear *delay-and-sum* beamformer steered to the end-fire [20] steering direction is shown. In the figure also a phenomenon called *spatial aliasing* can be seen in the high frequencies. Spatial aliasing is similar as temporal aliasing, where during sampling, the frequency components over Nyquist frequency, i.e. half sampling rate, are folded back to the desired frequency band. Similarly, spatial aliasing occurs if the inter-sensor spacing is larger than half wavelength of the highest frequency in the beamformer design bandwidth.

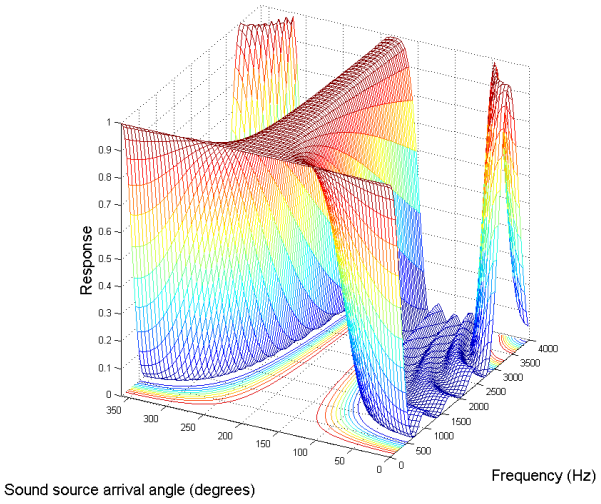


Figure 3.2: Response of a delay-and-sum beamformer.

In order to improve the DSB directivity and the frequency independence of the beam pattern, various methods have been suggested. For example a microphone array can be created by nesting several equally-spaced linear subarrays with different inter-sensor spacings, and using each subarray only for desired narrow-band frequency range. With this so called *harmonic nesting* constant beamwidth for wide bandwidth can be achieved, by combining the output of each subarray. This approach has been used e.g in [28] and in [19], where FIR filters are also used to control the frequency variations inside each subarray. By using unequally-spaced array and FIR low-pass filters a constant beamwidth for entire spatial region can be created as demonstrated by Ward *et al.* in [81, 82]. In their microphone array and beamformer design method, Ward *et al.* use low-pass filters to control the effective aperture size of an unequally-spaced linear array.

3.1.3 Beamformer Evaluation

To compare different beamformer designs, different performance measures have been developed. These measures include *array gain*, *beampattern*, *directivity index* and *output power* [11].

The most obvious reason for using an array of microphones instead of just one microphone is to improve the signal-to-noise ratio (SNR) of the output signal. This quality can be measured with *array gain* that can be formulated as:

$$G = \frac{SNR_{array}}{SNR_{sensor}}, \tag{3.4}$$

where SNR_{array} refers to the signal-to-noise ratio at the output of the microphone array and SNR_{sensor} is the signal-to-noise ratio in the output of just one sensor.

Beampattern or the beamformer spatial-temporal transfer function gives the beamformer response to a wavefront coming from a specific direction and specific frequency. An example of beamformer beampattern can be seen in Figure 3.2. It also should be noted that, in general 3D setup where the beampattern depends on azimuth and elevation angles, the beampattern can not be plotted in one single plot.

Output power is the measure of - as the name suggests - the total output power of the beamformer. This can be obtained from the beampattern by summing the beampattern over all the frequencies.

Directivity index is used to measure how well the beamformer suppresses signals from other directions compared to the steering direction. If $|H(\omega, \theta, \phi)|^2$ is used to denote the power spectrum of the beamformer as the function of frequency ω and azimuth θ and elevation ϕ angles, the directivity of the beamformer is

$$DI(\omega) = 10 \log_{10} \left(\frac{|H(\omega, \theta, \phi)|^2}{\frac{1}{4\pi} \int_0^{\pi} \int_0^{2\pi} |H(\omega, \theta, \phi)|^2 \sin(\phi) d\theta d\phi} \right). \tag{3.5}$$

Other measures to analyze beamformer performance also exist but are not discussed here in further detail. For a short description of these measures, the reader is advised to take look at [11].

3.1.4 Polynomial Beamformer

Essentially Polynomial Beamformer (PBF) [40, 44] is a *filter-and-sum* beamformer with adjustable filter characteristics. In PBF the FIR filter coefficients are approximated by using a polynomial function basis. The PBF approximation method is based on the well known

Farrow structure [27], where Taylor polynomials are used for continuous delay control in variable digital delay line. In the following the PBF filter structure is shortly reviewed.

To steer a *filter-and-sum* beamformer to multiple desired look directions $\mathbf{D}_i$, where $i = [1, 2, \ldots, N_D]$ and $\mathbf{D}_i$ defines both azimuth and elevation angles, $M \times N_D$ FIR filters have to be optimized. Thus, the output of the beamformer (3.2) becomes

$$y(n, \mathbf{D}_i) = \sum_{m=1}^{M} \sum_{k=0}^{L-1} h_{m,k}(\mathbf{D}_i) x_m(n-k). \tag{3.6}$$

In PBF the FIR filters are approximated as polynomial functions of order P

$$h_{m,k}(\mathbf{D}_i) = F_0(\mathbf{D}_i)a_0(m,k) + F_1(\mathbf{D}_i)a_1(m,k) + \ldots + F_P(\mathbf{D}_i)a_P(m,k), \tag{3.7}$$

where F_P is a scalar function of the continuous steering parameter $\mathbf{D}_i$. When F_P is chosen to be a Taylor polynomial, equation (3.7) is equivalent with the Farrow filter structure.

By inserting $h_{m,k}(\mathbf{D}_i)$ from equation (3.7) to (3.6), the output of PBF can be written as

$$y(n, \mathbf{D}_i) = \sum_{m=1}^{M} \sum_{k=0}^{L-1} \sum_{p=0}^{P} F_p(\mathbf{D}_i) a_p(m,k) x_m(n-k), \tag{3.8}$$

and by reordering the terms, the output of the PBF comes

$$y(n, \mathbf{D}_i) = \sum_{p=0}^{P} F_p(\mathbf{D}_i) \sum_{m=1}^{M} \sum_{k=0}^{L-1} a_p(m,k) x_m(n-k). \tag{3.9}$$

From (3.9) it can be seen that the PBF filter structure can be divided into two parts: the pre-filter and the post-filter. The pre-filter can be defined as

$$y'_p(n) = \sum_{m=1}^{M} \sum_{k=0}^{L-1} a_p(m,k) x_m(n-k) \tag{3.10}$$

and it creates P fixed intermediate beams. The post-filter

$$y(n, \mathbf{D}_i) = \sum_{p=0}^{P} F_p(\mathbf{D}_i) y'_p(n) \tag{3.11}$$

is then used to dynamically steer the intermediate beams to the desired look directions as shown in Figure 3.3. Because the post-filter calculations can be done very efficiently the PBF structure enables an easy way to create a desired number of parallel beams by controlling just one (2D) or two (3D) parameters. More detailed analysis of the computational performance and also a comparison with traditional *filter-and-sum* beamformer can be found in [44] and a PBF implemented as a GSC beamformer structure is presented in [56].

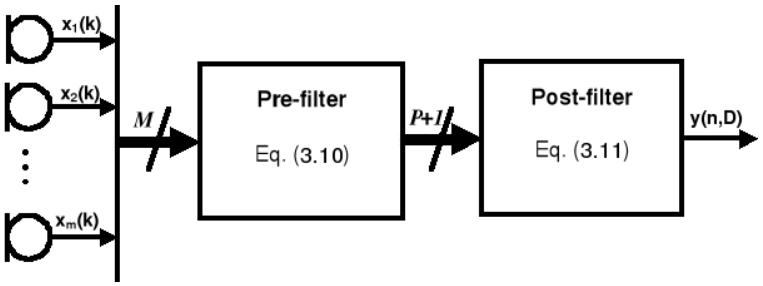


Figure 3.3: The Polynomial Beamformer (PBF) structure

The optimization of the pre-filter coefficients $a_p(m,k)$ is done by defining $\mathbf{\Omega}_d$ desired source signal locations and $\mathbf{\Omega}_n$ interference source locations. Next, $\mathbf{\Omega}_S = \mathbf{\Omega}_d \cup \mathbf{\Omega}_n$ defines all the source positions and $\mathbf{\Omega}_\phi$ defines the beamformer look directions. Now the Mean Square Error (MSE) between desired PBF output response and the actual response is defined as

$$e_{MSE} = \sum_{S \in \mathbf{\Omega}_s} \sum_{f \in \mathbf{\Omega}_f} \sum_{D \in \mathbf{\Omega}_\phi} \frac{(|\hat{Y}(z,D)| - |Y(z,D)|)^2}{|\mathbf{\Omega}_S| \cdot |\mathbf{\Omega}_f| \cdot |\mathbf{\Omega}_\phi|}, \tag{3.12}$$

where $\mathbf{\Omega}_f$ defines the frequencies and $Y(z,D)$ and $\hat{Y}(z,D)$ are the beamformer outputs in the frequency domain for the desired output response and the actual output response respectively. The pre-filter coefficients are optimized by minimizing the MSE function [43, 44].

3.2 Direction of Arrival Estimation

The goal of sound source localization algorithms is to estimate the source position of the incident sound wave captured by a microphone array. Most common sound source localization methods are based on the Time Difference of Arrival (TDOA) of the sound wave although other methods also exist. In most of the TDOA algorithms the delay between a reference microphone and another microphone in the array is estimated by using Generalized Cross Correlation (GCC), where an inverse Fourier transform of a weighted cross correlation spectrum between microphone pairs is calculated [17]. Various techniques have been proposed to estimate the sound source location from the TDOA estimates [18, 38, 84]

However, the sound source location can not only be estimated by TDOA but also by measuring the signal amplitudes or energies. Usually the energy-based sound source localization methods are based on the well known phenomenon that the intensity of a acoustic wave is inversely proportional to the squared distance. Some localization techniques that use this phenomenon are presented in [10, 69].

This Chapter starts with a theoretical discussion of the speech source localization framework in Chapter 3.2.1. In Chapter 3.2.2 a noise power estimation method is presented that is necessary when the direction of the beam with the largest SNR is desired. Finally, in Chapter 3.2.3 the creation of Spatial Likelihood Function (SLF) from the framework data is described.

3.2.1 Beamformer Based Speaker Localization Framework

In this work the sound source location is estimated by using steerable beamforming. A a similar approach as was proposed by Kellermann in [45]. The presented localization algorithm is based on [77] and it uses the PBF structure to generate a set of beams from which the beam with the largest signal power is defined. The Direction Of Arrival (DOA) of the sound source is then estimated to be in the steering direction of that beam. Finally, also the range of the sound source can be estimated by fusing the multiple DOA estimates produced by multiple microphone arrays.

When using a steerable beamformer the straightforward manner to estimate the dominant speaker direction is to find the beam with the largest signal power. However, especially in low Signal-to-Noise Ratio (SNR) conditions, for example inside a car, it is more beneficial to find the beam with the largest SNR than the beam with largest signal power. But in general the beam with largest SNR is not steered to the direction of active speaker. For example if the noise source is in the same direction as the speaker, it is not likely that the beam steered to that direction has the largest SNR. As the proposed system is intended for office use where generally the SNR is high, it is more likely that the speaker is at the steering direction of the beam with highest signal power. However, in the following both cases are considered.

As in the previous chapter let us consider that a beamformer output consists of N_D beams steered to (θ_i, ϕ_i) directions, where $i = [1,\ 2,\ \ldots\ , N_D]$. Now the output of the i^{th} beam of the PBF can be written as

$$\begin{aligned} y(n, \mathbf{D}_i) &= \sum_{p=0}^{P} F_p(\mathbf{D}_i) \sum_{m=1}^{M} \sum_{k=0}^{L-1} [a_p(m,k) s_m(n-k) + a_p(m,k) \vartheta_m(n-k)] \\ &= \sum_{p=0}^{P} F_p(\mathbf{D}_i) \sum_{m=1}^{M} \sum_{k=0}^{L-1} a_p(m,k) s_m(n-k) \\ &+ \sum_{p=0}^{P} F_p(\mathbf{D}_i) \sum_{m=1}^{M} \sum_{k=0}^{L-1} a_p(m,k) v_m(n-k) \\ &= y_s(n, \mathbf{D}_i) + y_\vartheta(n, \mathbf{D}_i). \end{aligned} \tag{3.13}$$

From equation (3.13) it can be seen that the source signal and the noise signal still have

a simple additive relation and are uncorrelated in each beam if $s_m(n)$ and $\vartheta_m(n)$ also are uncorrelated.

When the speech and noise signals are uncorrelated the total signal power of the i^{th} beam becomes

$$\begin{aligned} W_s(n,\mathbf{D}_i) = E\{|y(n,\mathbf{D}_i)|^2\} &= E\{|y_s(n,\mathbf{D}_i) + y_\vartheta(n,\mathbf{D}_i)|^2\} \\ &= E\{y_s(n,\mathbf{D}_i)^2\} + \underbrace{2E\{y_s(n,\mathbf{D}_i)y_\vartheta(n,\mathbf{D}_i)\}}_{=0} + E\{y_\vartheta(n,\mathbf{D}_i)^2\} \\ &= E\{y_s(n,\mathbf{D}_i)^2\} + E\{y_\vartheta(n,\mathbf{D}_i)^2\}. \end{aligned} \tag{3.14}$$

When the signal power estimate of the i^{th} beam in equation (3.14) is divided by the noise power, we have

$$W_r(n,\mathbf{D}_i) = \frac{E\{|y(n,\mathbf{D}_i)|^2\}}{E\{y_\vartheta(n,\mathbf{D}_i)^2\}} = \frac{E\{y_s(n,\mathbf{D}_i)^2\}}{E\{y_\vartheta(n,\mathbf{D}_i)^2\}} + 1. \tag{3.15}$$

As can be seen from equation (3.15), $W_r(n,\mathbf{D}_i)$ is proportional to the SNR of the i^{th} beam, while $W_s(n,\mathbf{D}_i)$ gives the total output power of the i^{th} beam. Now a choice can be made as to what quantity will be used for speech source DOA estimation. In general the choice depends on the acoustical environment and the actual application where speaker localization is utilized but based on our experimentation it can be said that if the microphone output signal SNR is low, then W_r will give better results than W_s, which leads to *beam level* estimate,

$$W(n,\mathbf{D}_i) = \begin{cases} W_s(n,\mathbf{D}_i), & \text{if } SNR > \delta \text{ dB} \\ W_r(n,\mathbf{D}_i), & \text{otherwise} \end{cases} \tag{3.16}$$

where δ is a SNR threshold value. However additional work has to be done in order to determine the desired value of δ.

Also, to make the DOA estimation more robust against rapid changes in acoustic power the beam level estimate is averaged over time instead of calculating the instantaneous beam level. The averaging is done by smoothing the beam level estimate with a first-order recursive system

$$P(n,\mathbf{D}_i) = \beta P(n-1,\mathbf{D}_i) + (1-\beta)W(n,\mathbf{D}_i). \tag{3.17}$$

For sufficient smoothing, the smoothing constant β should be set to values between $\beta = [0.90...0.95]$ and a step size between $[30\text{ms} - 70\text{ms}]$ should be used.

This information can be used in finding the DOA of speech signal by calculating $P(n,\mathbf{D}_i)$ of each beam and finding its largest value. However, since the absolute value of the beam level is not necessary for the localization algorithm and to make the previous beam levels

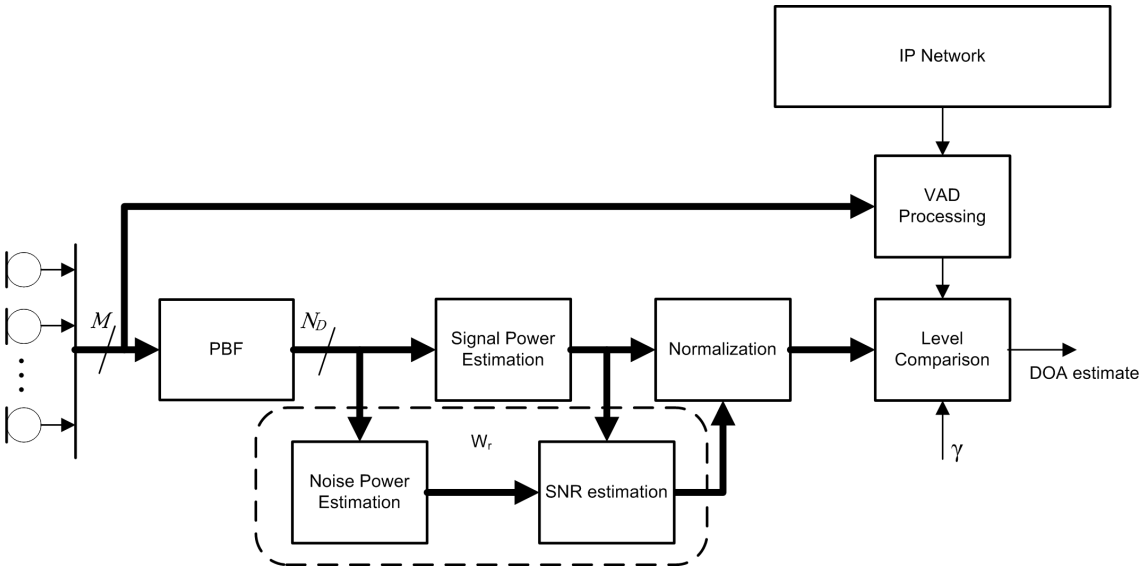


Figure 3.4: The speech tracker

also comparable the beam level estimate of each beam is normalized with the average beam level. Now we have

$$P_{avg}(n, \mathbf{D}_i) = N_D \frac{P(n, \mathbf{D}_i)}{\sum_{i=0}^{N_D-1} P(n, \mathbf{D}_i)}. \tag{3.18}$$

Finally, the normalized beam level estimates $P_{avg}(n, \mathbf{D}_i)$ have to be compared in order to find the beam $\mathbf{B}$ with the highest beam level $\hat{P}(n)$

$$\mathbf{B} = \arg\max_{\mathbf{D}_i} \{P_{avg}(n, \mathbf{D}_i)\} \tag{3.19}$$

$$\hat{P}(n) = P_{avg}(n, \mathbf{B}) \tag{3.20}$$

and thus also the estimate for the speech source direction. However, when tracking a speaker in a two-way communication system the tracking should be done only during the near-end speech activity and the far-end speech pauses. Therefore a voice activity detector (VAD), such as [52], is needed for robust speaker tracking. In here the voice activities are assumed to be known.

Thus, during the far-end speech pause and near-end speech activity, first the beam with the highest normalized beam level $\hat{P}(n)$ is calculated. Then a decision of a new speaker direction is made by comparing the relative level difference between $P_{avg}(n, \mathbf{B})$ and $P_{avg}(n-1, \mathbf{B})$ to a threshold value γ. The new speaker direction is decided by,

$$\hat{\mathbf{D}}(n) = \begin{cases} \mathbf{B}, & \text{if } \frac{P_{avg}(n,\mathbf{B})}{P_{avg}(n-1,\mathbf{B})} > \gamma \\ \hat{\mathbf{D}}(n-1), & \text{otherwise} \end{cases} \tag{3.21}$$

The threshold value γ can be chosen freely but the choice depends on the overall SNR. In low SNR conditions the possible dynamic range is reduced, which suggests the use of lower values of γ. However, experience has show that values near 1 dB provide satisfactory results in high SNR conditions. A block diagram of the presented tracking system can be found from Figure 3.4.

3.2.2 Noise Power Estimation

When using W_r as the beam level estimate the critical part of the localization algorithm is the estimation of the noise power that is needed for successful estimation of the SNRs of each beam. The obvious way to estimate the noise power is to calculate the total signal power when speech is not present i.e during a speech pause. This approach however requires a VAD whose performance can be unpredictable especially in low SNR conditions or in non-stationary noise. Therefore during false speech pause detection the noise power estimate can have very large variance. As already stated before, the speaker localization requires a VAD for robust speaker localization, which suggests that the noise power estimation can also be done using the VAD approach. However, the robustness requirements for VAD in power comparison phase of the tracker and in the noise estimation deviate largely. The false activity detection in power comparison does not cause large deterioration in the tracker performance but this is not the case in noise power estimation.

Therefore an algorithm for noise power estimation that does not require VAD is presented here. The described noise power estimation method is based on the minimum statistics approach by Rainer Martin [50, 51]. In minimum statistics the noise power spectrum is estimated by tracking the minimum value of a smoothed power spectrum of the signal inside a sufficiently long search window. This minimum value gives a biased noise power estimate for each frequency bin. By analyzing the statistics of the minimum value noise power estimator, the mean of the estimate can be found and it can be used to produce an unbiased estimate of the noise power.

However, because in the speaker localization framework only the beam's SNR relative to SNR of other beams is of interest, the unbiased estimate of the noise power is not needed. Also, instead of the power spectrum, the noise power in the full frequency band is estimated. Before search for the minimum value the signal power estimate can also be smoothed similarly as in equation (3.17) and we get

$$E_s(n, \mathbf{D}_i) = \alpha(n, \mathbf{D}_i) E_s(n-1, \mathbf{D}_i) + [1 - \alpha(n, \mathbf{D}_i)] W_s(n, \mathbf{D}_i), \tag{3.22}$$

where $\alpha(n, \mathbf{D}_i)$ is an adaptive smoothing parameter. In the most trivial case $\alpha(n, \mathbf{D}_i)$ can be a constant. Finally, a biased noise power estimator can be formulated as

$$\hat{\sigma}_n^2(n, \mathbf{D}_i) = min\{E_s(n, \mathbf{D}_i), E_s(n-1, \mathbf{D}_i), \ \ldots, E_s(n-Q+1, \mathbf{D}_i)\}, \tag{3.23}$$

where Q defines the length of the search window.

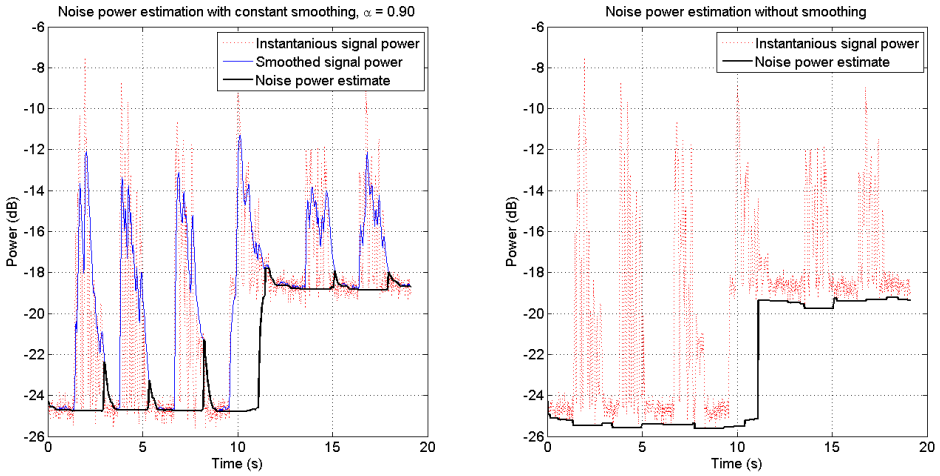


Figure 3.5: Noise power estimates. Red: instantaneous signal power, Blue: smoothed signal power $\alpha(n, \mathbf{D}_i) = 0.90$, Black: minimum tracked noise power estimate, search window length $Q = 96$.

In Figure 3.5 a minimum tracked noise power estimate is shown. In the left figure the noise power is estimated from the smoothed signal power and in the right figure there is no smoothing. As can be seen from Figure 3.5, the smoothing widens the power peaks during speech activity and therefore causes inaccurate noise power estimates. On the other hand if the smoothing is omitted from the signal power computation, also the inaccuracy due to the wide power peaks is avoided, but instead the variance of the noise power estimate is increased and the estimate is more biased as can be seen from the left figure. However, this can be avoided by using adaptive smoothing parameter as suggested in [51]. The optimal smoothing parameter can be derived by assuming a speech pause and minimizing the conditional mean square error

$$N_{MSE}(n, \mathbf{D}_i) = min\left\{E\left[(E_s(n, \mathbf{D}_i) - \sigma_n^2)^2 | E_s(n-1, \mathbf{D}_i)\right]\right\}, \tag{3.24}$$

where E is the expectation operator.

This leads to the optimal smoothing parameter

$$\alpha_{opt}(n, \mathbf{D}_i) = \frac{1}{1 + \left(\frac{E_s(n-1, \mathbf{D}_i)}{\sigma_n^2(n, \mathbf{D}_i)} - 1\right)^2} \tag{3.25}$$

As can be seen from equation (3.25) the smoothing parameter should be adapted so that the value of the parameter is close to one during speech pause and close to zero during

speech activity. The detailed derivation of the optimal smoothing parameter as well as its practical estimation can be found in [51]. Minimum tracker noise power estimate with adaptive smoothing parameter is shown in Figure 3.6.

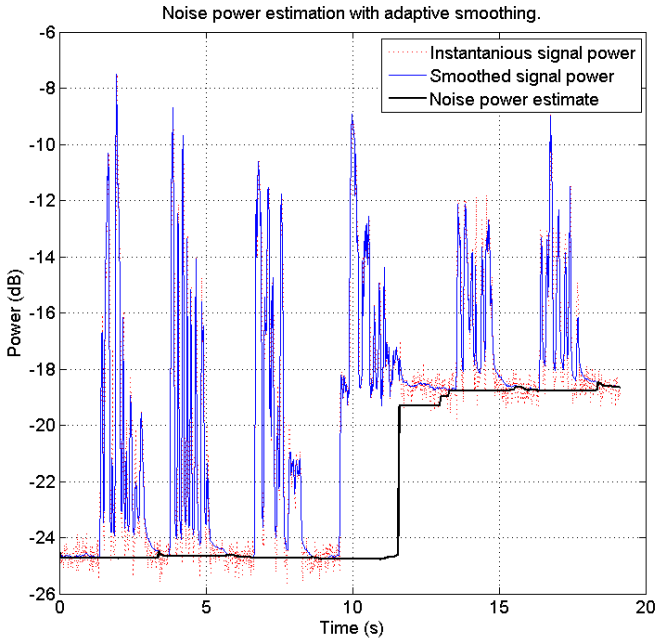


Figure 3.6: Noise power estimation with adaptive smoothing parameter

3.2.3 Spatial Likelihood Function

As described in Chapter 3.2.1, speaker localization framework gives an estimate for the most likely speaker direction, but in many cases it is also useful to analyze the likelihood of other speaker directions [5]. This is especially useful when fusing data from several individual estimations. The Spatial Likelihood Function (SLF) [5] or Spatial Response essentially describes a probability $Pr(\phi(\mathbf{p})|\mathbf{x})$, where $\phi(\mathbf{p})$ is the event that the speaker is located at point $\mathbf{p}$ and $\mathbf{x}$ is a vector containing the input data of all the microphones. Therefore $Pr(\phi(\mathbf{p})|\mathbf{x})$ can be thought as a *a posteriori* probability for a sound source to locate in point $\mathbf{p}$ in space. However, generally the absolute value of SLF can be difficult to measure and therefore there exist methods for measuring

$$SLF(\mathbf{p}) = (\psi(Pr(\mathbf{p})), \tag{3.26}$$

where $\psi(\cdot)$ is a monotonically non-decreasing function. Using $\psi(\cdot)$ does not impose further problems, because only the relative values of SLF at different source directions or locations

are of interest and using monotonically non-decreasing function does not violate the comparison of these values. In this work whenever likelihood or SLF is referred it is related to likelihood as defined in equation (3.26).

For the energy-based localization framework the SLF is produced by mapping the values of $P_{avg}(n, \mathbf{D}_i)$ to the corresponding directions. This can be formulated so that first the azimuth and elevation angles for each possible source location $\mathbf{p}$ relative to array at $\mathbf{p}_m$ are calculated as

$$\begin{aligned} \theta_{x,y,z} &= \arctan\left[\frac{p_y - p_{m,y}}{p_x - p_{m,x}}\right] \\ \phi_{x,y,z} &= \arctan\left[\frac{p_z - p_{m,z}}{\sqrt{(p_x - p_{m,x})^2 + (p_y - p_{m,y})^2}}\right]. \end{aligned} \tag{3.27}$$

Next, all possible source points $\mathbf{p}_s$ in space that are included in the area covered by steering direction $(\mathbf{D}_i)$ are given a value $P_{avg}(n, \mathbf{D}_i)$, i.e.

$$SLF(n, \mathbf{p}) = \begin{cases} P_{avg}(n, \mathbf{D}_i), & \text{if} \begin{cases} \theta_{x,y,z} - \Delta\theta \leq \theta_i < \phi_{x,y,z} + \Delta\theta & \bigwedge \\ \phi_{x,y,z} - \Delta\phi \leq \phi_i < \phi_{x,y,z} + \Delta\phi \end{cases} \\ 0, & \text{otherwise} \end{cases} \tag{3.28}$$

where $(\theta_i,\ \phi_i)$ are the steering angles of the i^{th} beam and $\Delta\theta$, $\Delta\phi$ are tolerance parameters that define the beam width in SLF calculations. For example when the steering directions are equally spaced in azimuth plane, the tolerance parameter can be calculated as: $2\Delta\theta = 2\pi/N_\theta$, where N_θ is the number of beams in the azimuth plane. However, it is not obvious that $SLF(n, \mathbf{p})$ can be used as an estimate for the likelihood function but this can be proven similarly as cross-correlation based SLF is proven in [5].

Essentially $SLF(n, \mathbf{p})$ is an observational estimate of $Pr(\mathbf{x}|\phi(\mathbf{p}))$, which describes the probability of the input data to be $\mathbf{x}$, in case of an event that the speaker is located at point $\mathbf{p}$. The relation between $Pr(\mathbf{x}|\phi(\mathbf{p}))$ and $Pr(\phi(\mathbf{p})|\mathbf{x})$ is found with with Bayes' theorem which states that

$$Pr(\phi(\mathbf{p})|\mathbf{x}) = \frac{Pr(\mathbf{x}|\phi(\mathbf{p}))Pr(\phi(\mathbf{p}))}{Pr(\mathbf{x})} \tag{3.29}$$

From equation (3.29) it can be seen that $Pr(\mathbf{x})$ is not dependent on $\mathbf{p}$ and therefore does not change the relative values of SLF at different points in space. Also, $\phi(\mathbf{p})$ describes the *a priori* probability for the event that the speaker is located at point $\mathbf{p}$. If it is assumed that all the possible points in space are possible source locations, $\phi(\mathbf{p})$ is constant over $\mathbf{p}$, which makes the SLF only a constant scaling of the observational estimate $Pr(\mathbf{x}|\phi(\mathbf{p}))$.

Figure 3.7 shows an example of an SLF that is produced with tracking framework described in Chapter 3.2.1. The red line marks the 0° direction and the angle increases clockwise. Larger values indicate higher speech source likelihood and thus the speaker is estimated to locate at the direction of 130°.

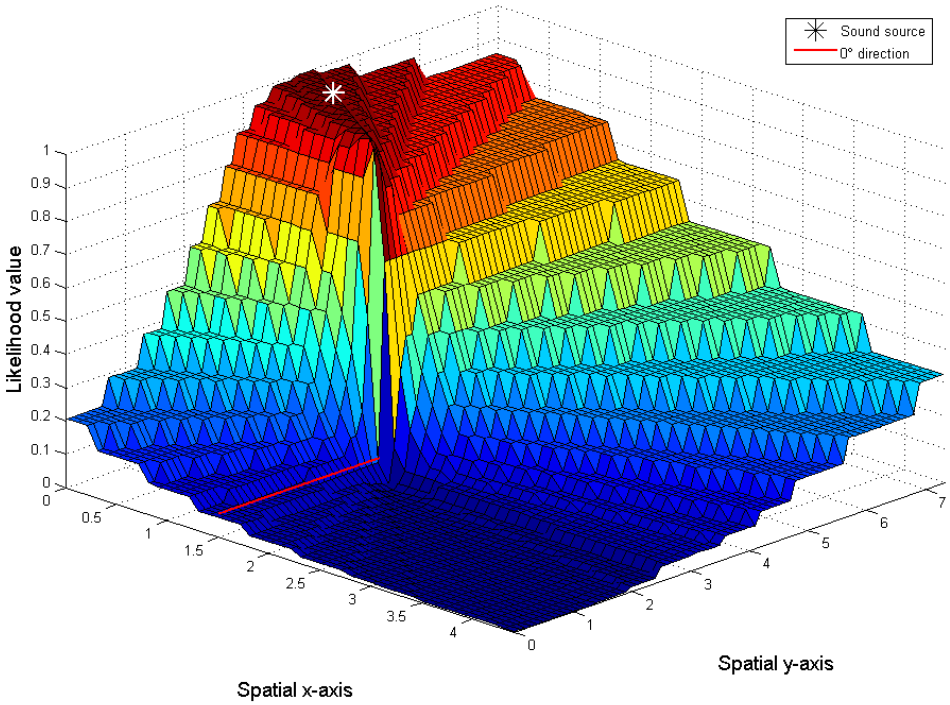


Figure 3.7: Example of a Spatial Likelihood Function, from recordings with circular microphone array of eight microphones [40].

3.3 Distributed Microphone Arrays

In this Chapter few possibilities of distributed signal processing are discussed. In Chapter 3.3.1 a short introduction and motivation for distributed signal processing is given. The fusing of the speech source DOA data of each microphone array in order to obtain the speech source location estimate is described in Chapter 3.3.2.

3.3.1 Background

As already stated, in beamforming the array sensitivity is increased to the direction of the desired source, while at the same time noise sources from other directions are attenuated. However, the attenuation of signal energy due to the distance that the acoustic wave travels, decreases the SNR of the microphone array and therefore also the parameter estimation error increases. An obvious solution for this is to place microphone arrays at different points in a room, thus decreasing the effect of attenuated signal energy. Also, by combining the DOA data of each microphone array the direction and/or position of sound source can be estimated more accurately.

In distributed signal processing each processing unit, e.g. one microphone array, is called a *node*. Node is responsible for all the local signal capture and processing. The communication between individual nodes depends of the topology of the sensor network which in theory can change very frequently [7]. One of the most common topologies is so called star topology, where each node sends its data to a *master node* that combines the data from the individual nodes and, when necessary, sends the combined data forward. The main problems in distributed audio signal processing, where the sensor topology can be assumed to be constant, are the transfer and the fusion of the data of each processing node. In this work transferring the data from node to node is assumed to be perfect, i.e. the internal clocks of each node are synchronized with each other, there are no latencies in the transfer, etc.

3.3.2 Sound Source Localization from Multiple DOA Estimates

The fusion of the individual microphone array data can be done in several ways. In [48] Liu *et al.* estimate the sound source position by finding the intersection of several DOA lines from each microphone array. In the case of non-intersecting lines, the sound source position is defined to be the point that has minimal overall distance from the lines. In ideal conditions this method gives optimal sound source position estimate, but errors due to noise sources, inaccurate microphone calibration, etc. can lead to erroneous DOA estimates and therefore also erroneous sound source position estimate.

Figure 3.8 presents an example of a simulated setup of four microphone arrays, where each microphone array consists of 10 ideal cardioid microphones, thus creating cardioid

shaped beams in 36° intervals. The DOA estimates for each microphone array are formed by using the localization framework presented in Chapter 3.2.1 and finally the DOA estimates are combined by searching the minimal overall distance from the DOA lines. The left figure presents conditions with ideal microphone calibration whereas in the right figure the microphone gains are uniformly distributed between ± 0.5 dB. The input signal is English female speaker. The blue square shows the real speaker position and black star shows the estimated speaker position. The microphone array positions are marked with red circles.

The fusion of DOA data can also be done by using the Spatial Likelihood Function presented in Section 3.2.3. This way also the likelihood of other DOA estimates is taken into account in the sound source position estimation. By using not only the most likely DOA estimate, but also the likelihood of other DOA estimates, the sound source localization can be made more robust against estimation errors.

The most straightforward way for combining the SLFs of each array is to sum the SLFs of each individual array to form a global estimate of the SLF function as shown in [5]. Now a global SLF is

$$\overline{SLF}(n, \mathbf{p}) = \frac{1}{N_M} \sum_{i=1}^{N_M} SLF_i(n, \mathbf{p}), \tag{3.30}$$

where N_M is the number of microphone arrays and $SLF_i(n, \mathbf{p})$ is the SLF generated by the i^{th} array. Now the speaker position estimate can found by searching a maximum of $\overline{SLF}(n, \mathbf{p})$, i.e.,

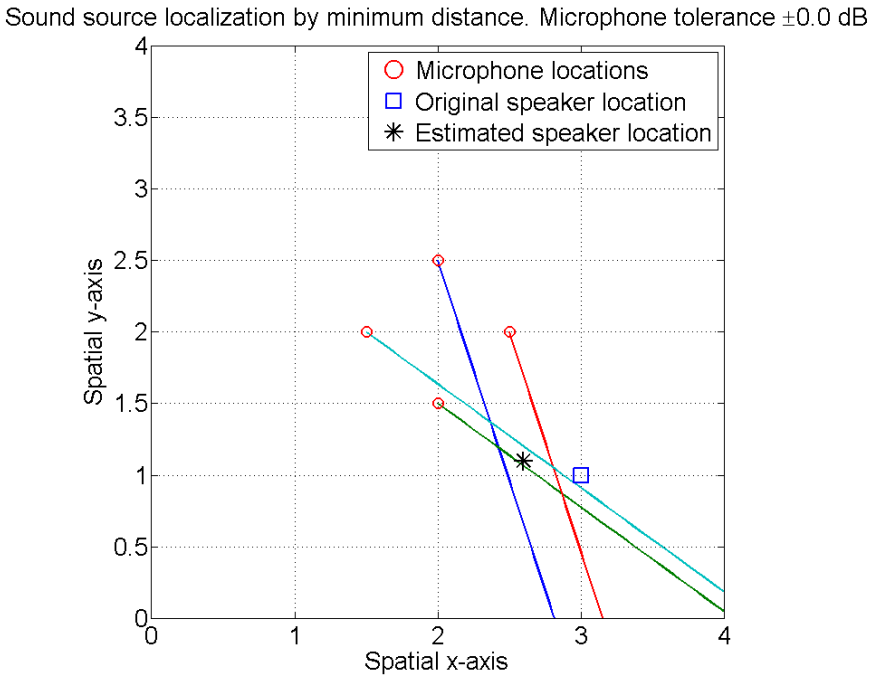
$$\hat{\mathbf{p}}(n) = \arg\max_{\mathbf{p}} \{\overline{SLF}(n, \mathbf{p})\} \tag{3.31}$$

However, when SLF is created using the localization framework as presented in Section 3.2.1, it should be noted that in general there are several source points with the same maximum value of $\overline{SLF}(n, \mathbf{p})$. This is due to the finite DOA estimate resolution which creates areas with the same SLF values. The final speaker position estimate can be found as a midpoint of all the points with maximum SLF value

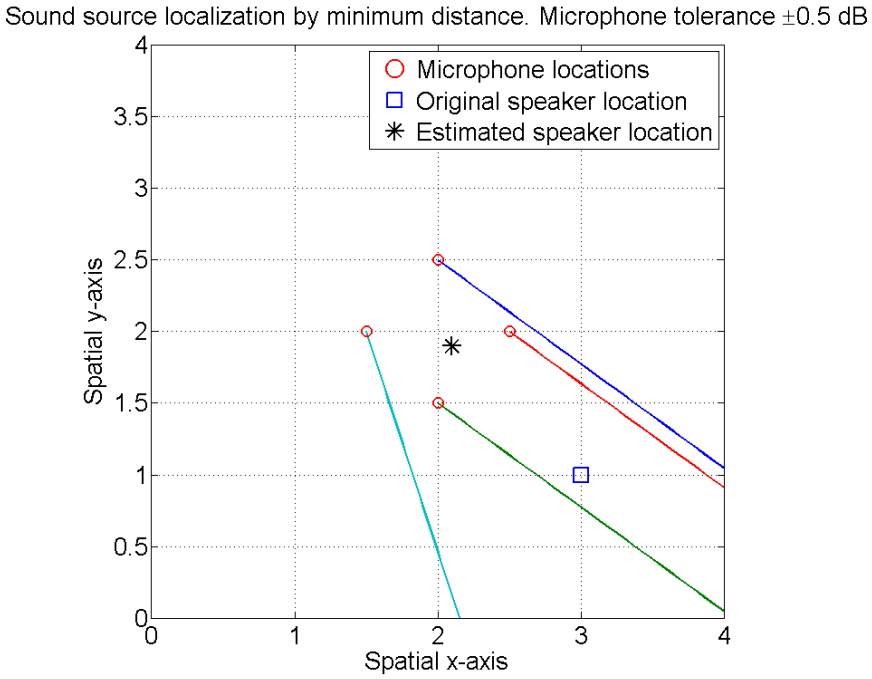
$$\hat{\mathbf{p}}_s(n) = \frac{1}{N_P} \sum_{i=0}^{N_P - 1} \hat{\mathbf{p}}_i(n), \tag{3.32}$$

where N_P is the number of points with the same maximum SLF value and $\hat{\mathbf{p}}_i$ is the coordinates of the i^{th} maximum value.

An example of global SLF can be seen in Figure 3.9, where the simulated setup is the same as in Figure 3.8. Here the sound source position is estimated by using the likelihood of all DOA estimates and thus forming a global SLF. Similarly as in Figure 3.8, the left figure shows the global SLF with ideal microphone calibration and in the right figure the microphone gains are uniformly distributed between ± 0.5 dB. Also as in Figure 3.8, the



(a) Ideal microphone calibration



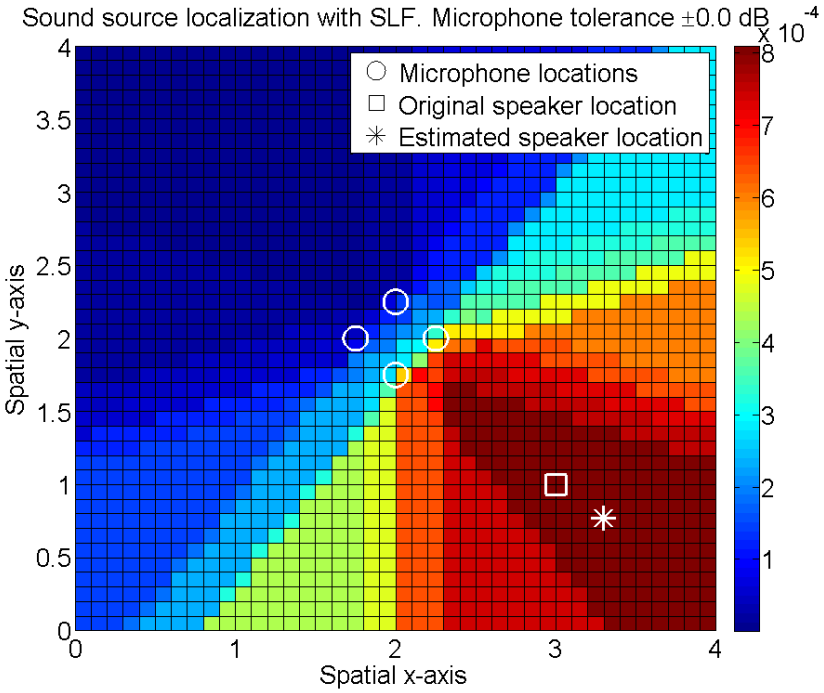
(b) Microphone gain ±0.5 dB

Figure 3.8: Sound source localization by searching the point with minimal overall distance from several DOA lines.

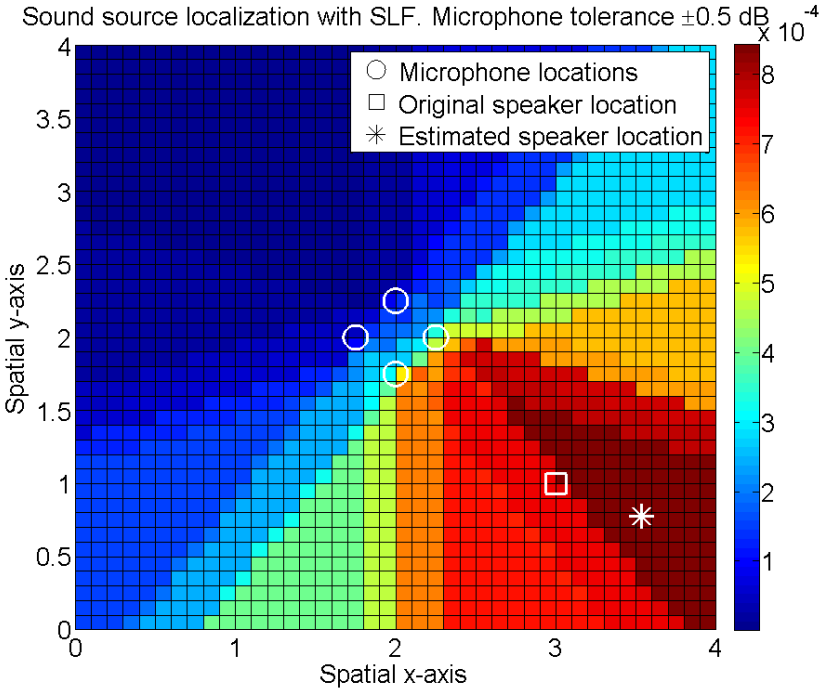
square shows the real speaker position, star shows the estimated speaker position and the microphone array positions are marked with circles.

As can be seen from Figures 3.8 and 3.9, in some situations the global SLF provides more robust speaker localization when there are errors in microphone calibration. The search of smallest distance can cause large speech source localization estimation errors, when even one of the individual DOA estimations deviates significantly from DOA estimates of other microphone arrays. In the case of global SLF this error is reduced by averaging over all the DOA estimates from all the microphone arrays. More thorough comparison of the two methods is presented in Chapter 4.4.2.

Also other methods for generating the global SLF have been proposed. A Maximum Likelihood (ML) estimate of the sound source location by using acoustic energy is presented in [69] and in [6] ML estimate is formulated by using relative time delays. A different kind of approach is presented in [60] where a kinematic model for speaker state (position and velocity) is generated and sequential Kalman filtering is used to fuse the state estimates of each microphone array.



(a) Ideal microphone calibration



(b) Microphone gain ±0.5 dB

Figure 3.9: Sound source localization by global SLF

Chapter 4

Experimentation

In the following chapters the implemented microphone array and its performance is discussed. First in Chapter 4.1, the design of the implemented microphone array and the motivation for the chosen design is discussed. Next, in the Chapter 4.2, the measurements to evaluate the system performance in normal acoustic conditions are presented and the next two Chapters 4.3 and 4.4 discuss the system performance in simulated and in measured conditions.

4.1 Microphone Array Design

4.1.1 Geometry

In this thesis the interest is in small microphone arrays and therefore it was decided to use just a few microphones for each microphone array. By restricting the number of microphones, also the hardware requirements can be held under control when using multiple microphone arrays. Four microphones was chosen to be an adequate number of microphones for each array.

For this work a circular microphone array with one microphone in the center was chosen. This way a microphone array with microphones placed in corners of equilateral triangle and in the circumcenter of the triangle is formed. The radius of the circle with center in the circumcenter of the triangle was chosen to be 35 mm, which makes the spatial aliasing limit to be approximately 5 kHz. The microphones were placed on top of a chipboard piece and the resulting microphone array can be seen in Figure 4.1 and the microphone locations are shown in Table 4.1. The final Distributed Microphone Array System (DMAS) presented in this work uses three of the aforementioned microphone arrays of four microphones. Three microphone arrays is sufficient to cover an area of a normal office room while also maintaining the total number of microphones in control.

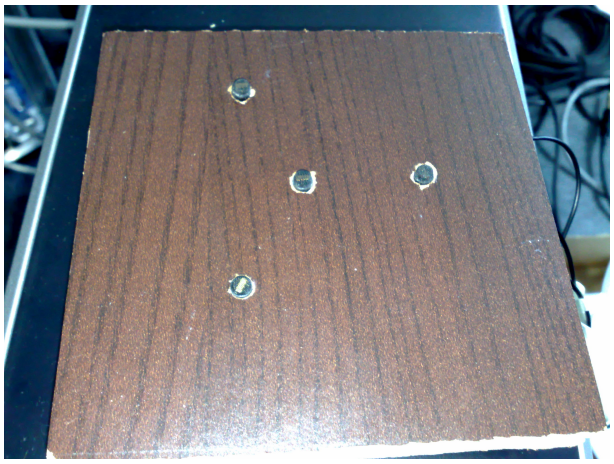


Figure 4.1: The prototype microphone array

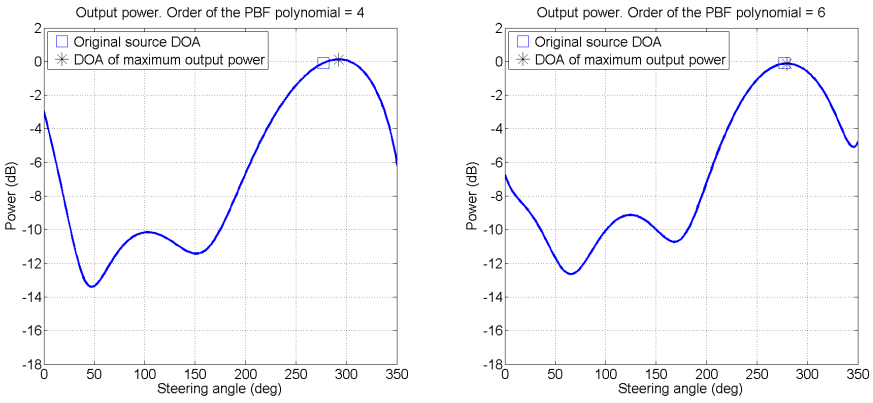
Mic	x[mm]	y[mm]	z[mm]
1	0	0	0
2	35	0	0
3	$35\cos(\frac{2\pi}{3})$	$35\sin(\frac{2\pi}{3})$	0
4	$35\cos(-\frac{2\pi}{3})$	$35\sin(-\frac{2\pi}{3})$	0

Table 4.1: Microphone locations in millimeters.

4.1.2 PBF Design

The PBF filter coefficients can be optimized for arbitrary shaped microphone arrays and with arbitrary beam steering properties. For this application the PBF was designed for a teleconference-like situation where the speakers are sitting around a table and the microphone arrays are placed on top of the table. Therefore it was decided to use fixed steering angle for elevation while maintaining continuous steering in azimuth plane. The PBF filter coefficients were optimized for elevation angles between 15° and 25°. In practice this means that if the speaker's mouth lies 0.5 m above the table surface, the optimum results are achieved when the microphone arrays is 1.07 m - 1.87 m from the speaker. The sampling rate of the PBF was chosen to be 8 kHz and the component tolerance was set to 0.5 dB.

When PBF is used for speaker localization, the most notable error source is the existence of a systematic bias error in the beamformer output power. This is demonstrated in Figure 4.2, where the output power of the PBF is plotted. As can be seen from the Figure 4.2(a) the DOA of the maximum output power is deviated over 10° from the real source DOA. This is due to the low order of the PBF polynomial that is used to approximate the beamformer filter coefficients. In real world situations also microphone calibration errors can cause systematic error in the PBF output. In Figure 4.2(b) a higher order polynomial is used and the systematic error is reduced significantly. In the figures black star shows the DOA with the highest output power and blue square shows the real DOA.



(a) Order of the PBF polynomial P = 4 (b) Order of the PBF polynomial P = 6

Figure 4.2: The bias error in the PBF output power.

When only one microphone array is used, the bias error does not effect the PBF performance, because the input signal is always filtered with the beam that provides the largest output power. However, when biased PBF outputs from multiple microphone arrays that

can have different bias errors in different directions are used to track a sound source, the bias errors are also present in the output of the whole distributed microphone array system. This deteriorates the performance of the whole system and therefore it has to be made sure that sufficiently high polynomials are used and that the microphones are calibrated. Therefore in this work, polynomials of order 6 are used to approximate the beamformer filter coefficients.

4.2 Measurements

A set of measurement were made to experiment and to test the distributed microphone array system and they are described in the following chapters. The measurements were made in collaboration with the Audio Research Group in the Department of Signal Processing in Tampere University of Technology (TUT).

4.2.1 Measurement Environment

Measurements were made in a listening room in the TUT audio laboratory. The room also included table, sofa and several audio equipment. The reverberation time T_{60} of the room was measured to be 0.25 s. Although this reverberation time is lower than in normal office or living room it is appropriate for testing and validating the system described in this work. The noise level of the room was below the sensitivity of B & K 2232 sound pressure meter.

4.2.2 Hardware

Commercial high quality audio hardware was used in the measurement. The microphones were required to be small enough to be suitable for microphone arrays and omnidirectional so that unwanted attenuation of signal energy does not happen due to the sound source location. The selected microphones were DPA 4060-BM pre-polarized omnidirectional miniature condenser microphones [1].

The microphones were connected to a Presonus Firepod microphone pre-amplifier and A/D converter [4], that was run with a desktop PC computer with Windows XP operating system. Cubase recording software was used to capture the microphone signals. The microphone signals were sampled using 48 kHz sampling frequency and 32 bits for each sample. Similarly Adobe Audition was used to play the source signals. The playback PC used RME-Hammerfall Multiface D/A converter that was connected to four Genelec 1029 A loudspeakers [2] for playback.

4.2.3 The System Geometry

In the measurement setup the microphone arrays were placed to simulate a teleconference-like situation, where all the microphones are on top of a table that is located between the speakers. Therefore, the microphone array locations were arbitraryly chosen but restricted to be on the table. The room layout and measurement geometry can be seen in Figure 4.3.

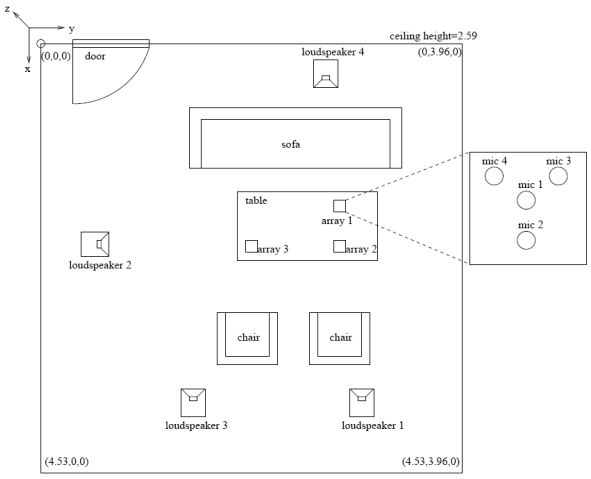


Figure 4.3: The recording setup.

4.2.4 Source Material

Several simulated conversational setups were created by using prerecorded speech samples from the TIMIT speech database. The speech samples include male and female speakers speaking in English. Also the impulse responses from each loudspeaker to each microphone were measured by using Farina's method [26].

4.2.5 Outcome of the Measurements

After the measurements some artifacts were found in the measured signals. Most prominently, as can be seen from Figure 4.4, a low frequency noise was present in the measurement data. The low frequency interference appears in all the measured data and is most likely due to the air conditioning system or other structural properties of the measurement room. However, the impulse response estimation method is able to push the interference out of the final impulse response estimate and therefore the impulse responses are not deteriorated by the low-frequency noise.

Also, the signals were measured with two different SNR levels. The source signal sound pressure was measured by playing white noise from the loudspeakers and measuring the

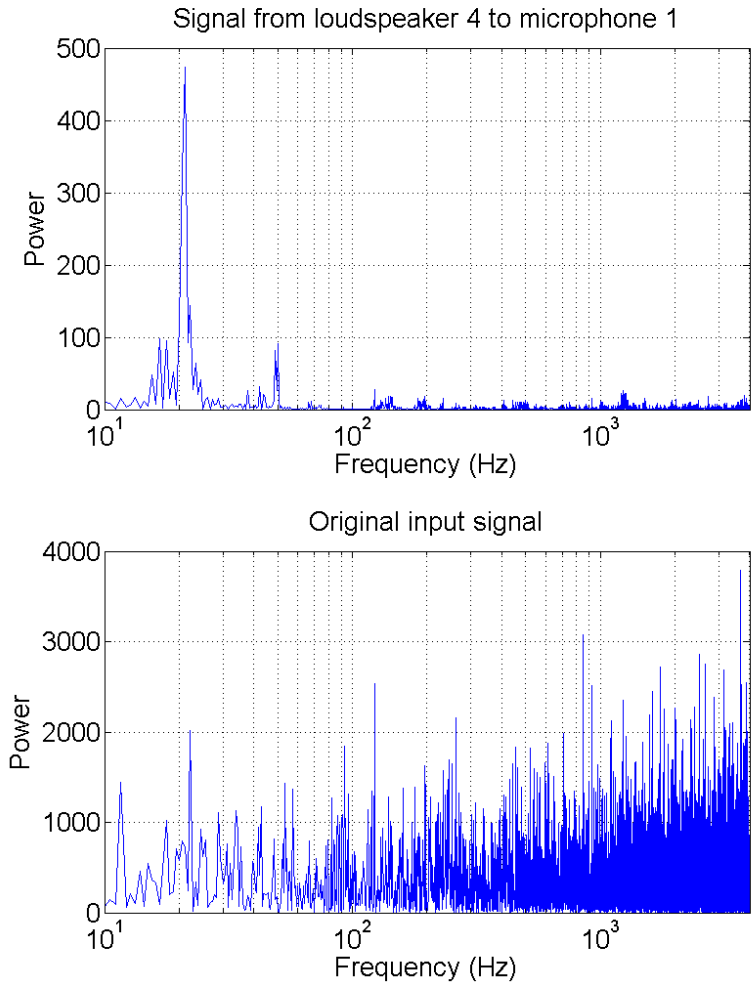


Figure 4.4: The power spectra of the output signal of the microphone 1 in array 1 and original Gaussian white noise input signal

sound pressure 30 cm from the loudspeaker. The SNR was modified by changing the output level of the loudspeaker. In high SNR conditions the sound pressure was approximately 81 dB and in low SNR conditions the sound pressure was approximately 30 dB lower. The sound pressure measurements were made A-weighted by using a B & K 2232 sound level meter.

No separate microphone calibration measurements were made and therefore the measured impulse responses were used for microphone calibration. The calibration was done by assuming that all the microphones of one array are placed at location of the array's center microphone and calculating the power of the impulse responses from all the loudspeakers to all the microphones. The impulse response's powers were averaged over all the loudspeakers and finally the gain coefficient for each microphone was defined so that the power of the impulse responses is the same at each microphone. However, in reality the microphones are not located at the same point in space and therefore the used calibration method is not ideal. The arrays are located approximately 1.5 m from the loudspeakers, which results in approximately 0.2 dB error in the microphone calibration, which is under the component tolerance used in the PBF design.

4.3 PBF Performance

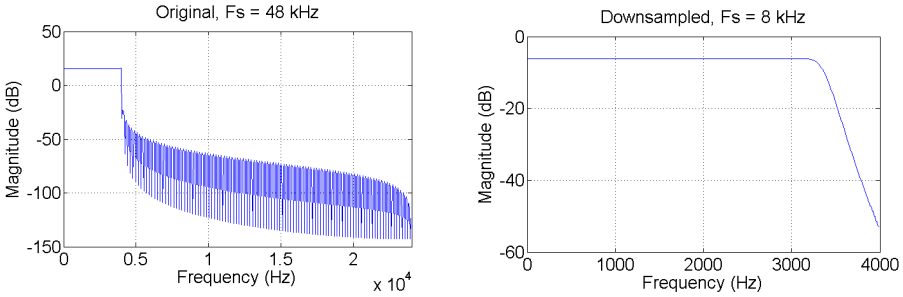
The implemented four-microphone Polynomial Beamformer performance is discussed in the following chapters. First the PBF performance in simulated setup is presented and finally performance in real acoustical environment is discussed.

4.3.1 Simulated PBF Performance

The PBF performance is first evaluated by simulating an anechoic environment. The geometry of the simulated setup is equivalent with the recording setup and also 48 kHz sampling rate was used in the simulations in order to make the results comparable. The simulations were made by using a synthesized baseband signal as the microphone input signal. However, the PBF was designed for 8 kHz sampling rate and therefore the microphone signals had to be downsampled before the PBF processing. The magnitude responses of the original and downsampled input signals can be seen in Figure 4.5.

When the source is placed at the location of the loudspeaker 4 and the microphone array 1 (as shown in Figure 4.3) is used to capture the signal, the beam pattern of the four microphone PBF and DSB can be seen in Figure 4.6.

As can be seen from the figure, the PBF design provides significantly more spatial attenuation than the DSB. The implemented array aperture is not sufficient for DSB to make separation of low frequency signals. The same behavior is also shown in Figure 4.7 where



(a) Synthesized input signal (b) Input signal after downsampling to 8 kHz

Figure 4.5: Magnitude responses of the input signal.

the directivity and the output power of the PBF and DSB are presented.

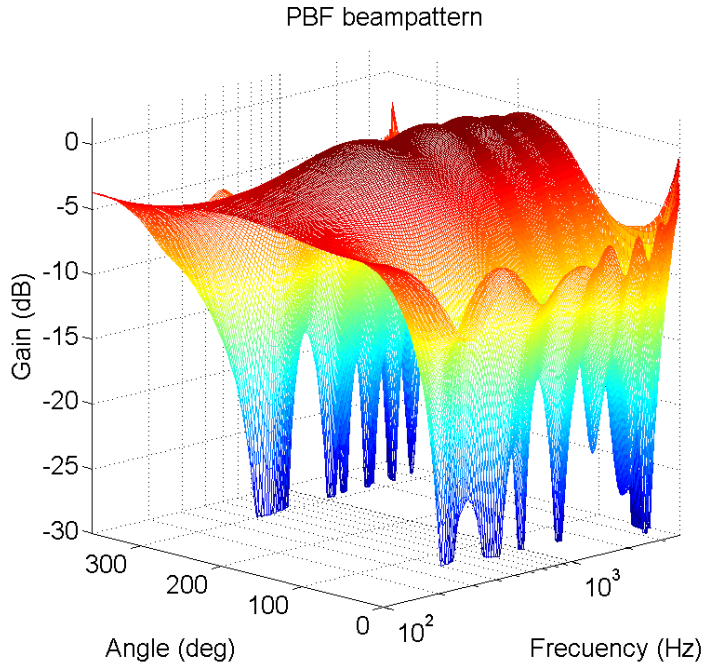
Figure 4.7(a) emphasizes what already was seen in Figure 4.6, that the PBF provides approximately 4 dB better directivity than DSB already at 500 Hz. Also in PBF the directivity is constant over wide frequency range. It should be noted that the strange behavior in directivity near 4 kHz is not because of imperfections in the beamformer design but is due to the imperfect anti-aliasing filtering during downsampling. The output powers in Figure 4.7(b) show that with PBF, a sidelobe attenuation of approximately 7 dB can be achieved whereas with DSB the sidelobe attenuation is only 3 dB. Also, PBF gives 10 dB more of maximum attenuation (difference between maximum and minimum) than DBF. The beam pattern and the output power are normalized with the output power of the microphone 1 that is located at the phase center of the array.

4.3.2 Measured PBF Performance

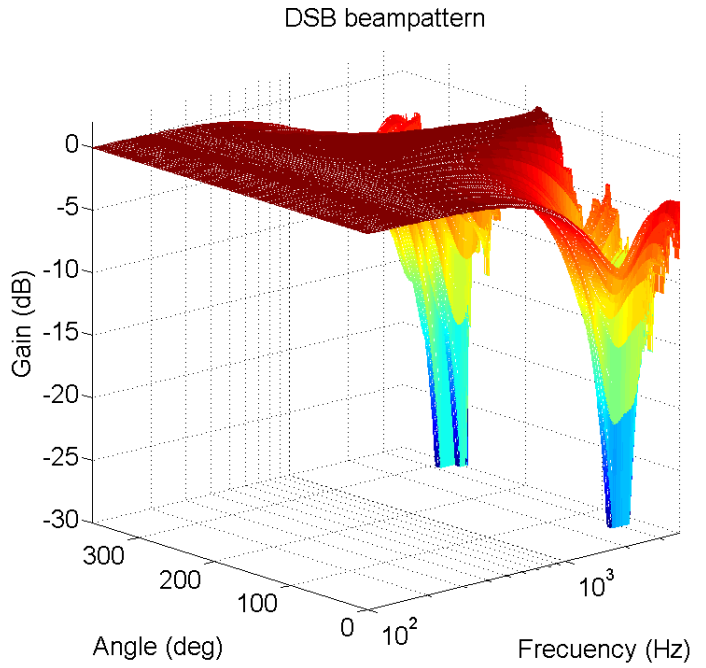
The PBF performance in real acoustic environment was evaluated by doing measurements that are described in Chapter 4.2. It has to be noted that most of the beamformer performance measures are not well-defined in multi-path propagation conditions and therefore they can not be used to evaluate the beamformer performance. However some observation can be made by looking at the directivity and total output power of the beamformer.

Figure 4.8 shows the total output power and the directivity of the PBF (blue) and DSB (red), when the source is played from loudspeaker 4 and captured with microphone array 1. The measured impulse response of the signal path from the loudspeaker to each microphone is used as the PBF input signal. The output powers are normalized with the output power of the microphone 1.

As can be seen from Figure 4.8(b), the PBF maximum attenuation is approximately 2.5 dB more than with DSB, but also PBF reduces the signal power almost 2 dB compared to

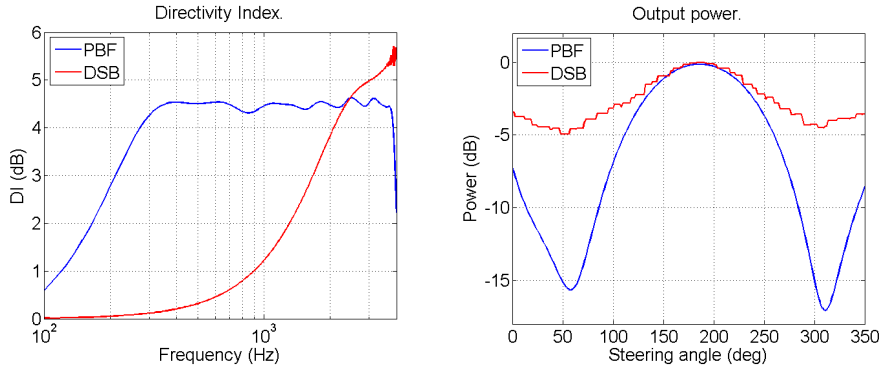


(a) PBF beam pattern, steered towards loudspeaker 4 from microphone array 1



(b) DSB beam pattern, steered towards loudspeaker 4 from microphone array 1

Figure 4.6: Beam patterns of the implemented microphone array used as PBF and DSB.



(a) PBF and DBF directivity index

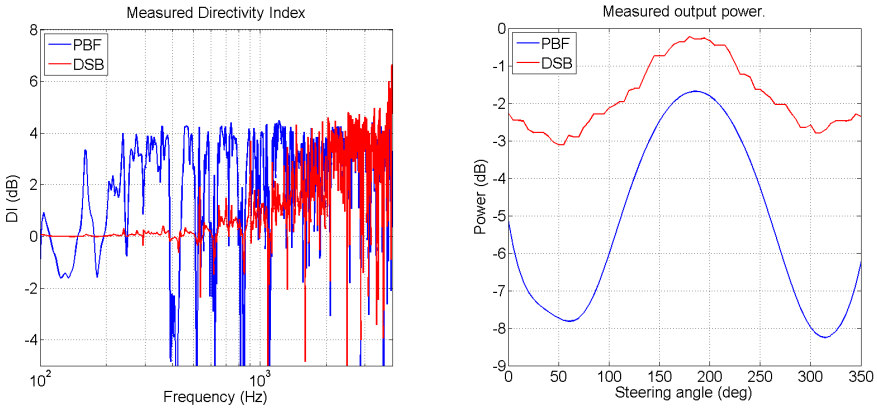
(b) PBF and DBF output power

Figure 4.7: Directivity and output power of PBF and DSB.

omnidirectional signal. Compared to the simulated setup, the dynamic range of the PBF output power is extensively reduced in real acoustic environment as can be expected.

The measured directivity in room environment in Figure 4.8(a) shows the same kind of behavior as in simulated case but now the directivity is deteriorated by the room properties. From the figure it can be noted that at low frequencies DSB can not separate the signals from different directions at all, while PBF has 3 dB directivity at 300 Hz. But at the same time large dips in the measured directivity can be seen at approximately 400, 500 and 600 Hz, which most likely are due to the room resonances in the measurement room. The outcome is that at these frequencies PBF output power is larger at other directions than the sound source direction.

However, when speech signal is used as the PBF input signal, i.e. the PBF output is weighted with speech spectrum, the PBF performance is increased when compared to DSB. This is demonstrated in Figure 4.9. As it can be seen from the figure, DSB does not attenuate speech signal arriving from different directions. This is because the small aperture of the microphone array causes the directivity of the DSB to be zero over wide frequency range as shown in Figure 4.7(a) while at the same time speech signals do not have much power in the high frequencies where DSB can discriminate with signals arriving from different directions.



(a) PBF and DBF measured directivity in multi-path propagating environment

(b) PBF and DBF measured output power

Figure 4.8: Measured output power and directivity of PBF and DSB.

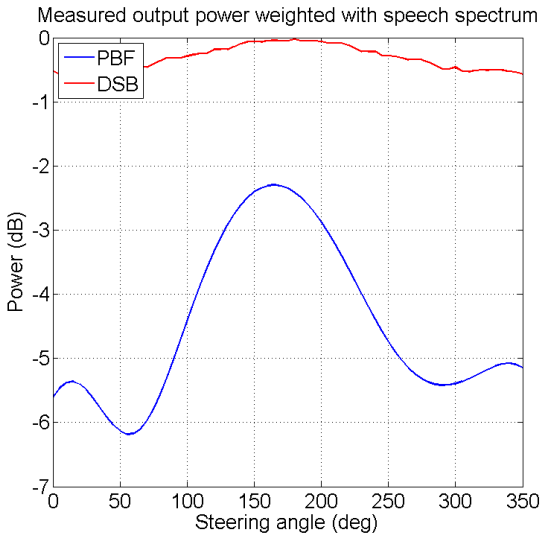


Figure 4.9: Measured output power of PBF and DSB weighted with speech spectrum.

4.4 Speech Source Localization Performance

Here the presented speech source localization framework is evaluated first by simulations and then with measured data.

4.4.1 Noise Power Estimation

When sound source localization is done in low SNR conditions it might be useful to use the beam with the best SNR as a speech source DOA estimate instead of the beam with the highest power. This may result in inaccurate DOA estimation but on the other hand when the DOA estimate is used to steer a beamformer the final beamformer output will have better SNR. When the desired signal and the noise are zero-mean and uncorrelated the SNR can be estimated with:

$$SNR = 10\log_{10}\left(\frac{E\{y(n)^2\}}{E\{\vartheta(n)^2\}} - 1\right), \tag{4.1}$$

where $y(n)$ is the desired signal contaminated with noise $\vartheta(n)$ and E is the expectation operator. As it can be seen from equation (4.1) the difficulty in SNR estimation is the estimation of the noise power.

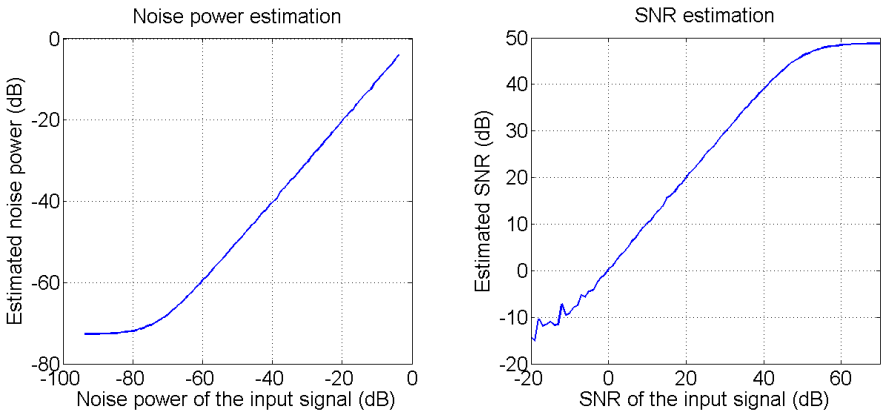
The minimum statistics noise power estimation method presented in Chapter 3.2.2 was evaluated by using a speech signal of English speaking female as a desired signal and adding white Gaussian noise to the desired signal. As the noise power estimation was done separately for each audio frame (length 256 samples) the final noise power estimate was calculated by averaging over the noise power estimates of all the frames. Figures 4.10(a) and 4.10(b) show the noise power estimate and the SNR estimate, respectively.

The figures show that at very low noise power levels the estimation saturates and therefore is not anymore reliable. However, between SNR -5 dB to 40 dB the presented noise power estimation method gives good results. Because the noise power estimation is not the main point of this work the evaluation or the method is done only for stationary white noise. With different type of noises the noise power estimator performs differently and therefore any generalizations can not be made. A more detailed evaluation of a similar noise power estimation method is given in [51].

4.4.2 Simulated Source Localization Performance

One array

In far-field the Direction Of Arrival (DOA) of the sound source can be estimated, when only one microphone array is used. Here, the DOA estimation method descried in Chapter 3.2 and Time Difference of Arrival (TDOA) based DOA estimation method are compared by



(a) The noise power estimation

(b) The SNR estimation

Figure 4.10: Noise power and SNR estimation

simulating anechoic environment in a noise field. The TDOA method used is described in [88], where the least squares solution of the DOA is calculated with estimated time delays. The time delays are estimated by using cross-correlation between the microphone signals and finding its maximum value.

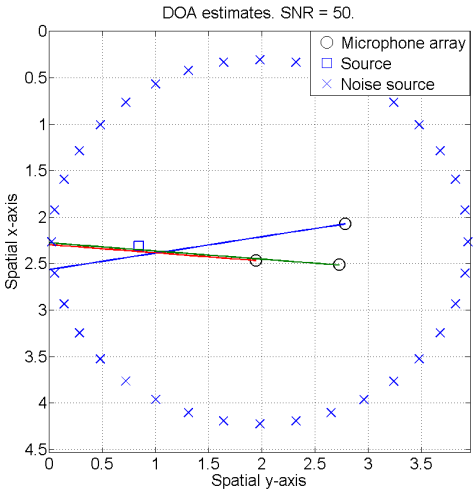
The first set of simulations are done by using a speech sample of English speaking female as the microphone input signal and placing uncorrelated Gaussian white noise sources at 10° intervals around the room shown in 4.3. All the signals are sampled at 48 kHz. The delayed microphone output signals are created by using linear interpolation between the samples[1]. The PBF is steered to 72 different directions which gives 5° as the DOA estimation resolution. The threshold value γ in equation (3.21) is set to 1 dB in high SNR conditions and to 0.2 dB at low SNR condition and the estimation is done with 0.5 s long audio frames. Total output power of the PBF is used as beam level estimate.

Figure 4.11 shows the DOA estimates from each microphone array to loudspeaker 2, when the SNR is 50 dB at one meter from the source and Figure 4.12 shows the DOA estimates when SNR is 0 dB.

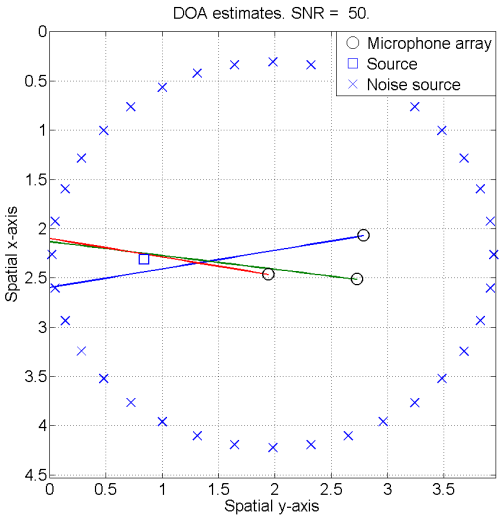
As can be seen from Figures 4.11 and 4.12, in this example the proposed method performs equally well in both high and low SNR conditions. The TDOA method produces slightly larger error in low SNR conditions than the proposed method. As expected the estimation error is larger especially at the arrays further away from the sound source.

In the second set of simulations only one noise source (in real environment this could be e.g. air conditioning) is placed at direction of 180° and distance of 2 meters from the

[1]MATLAB `interp1` function

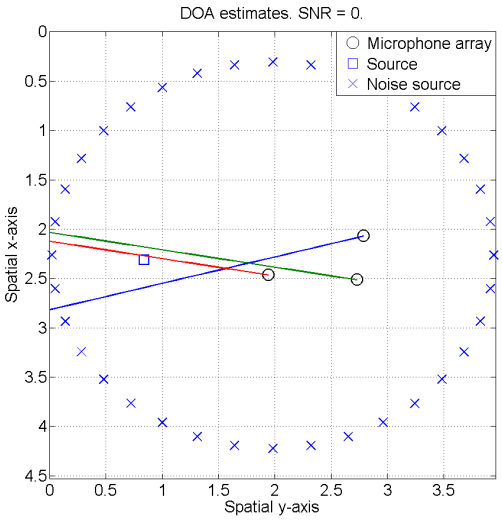


(a) The proposed DOA estimation.

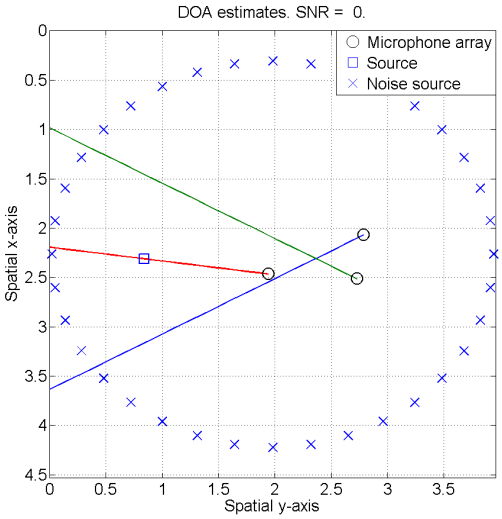


(b) TDOA based DOA estimation [88].

Figure 4.11: DOA estimation in noise field. SNR = 50 dB.



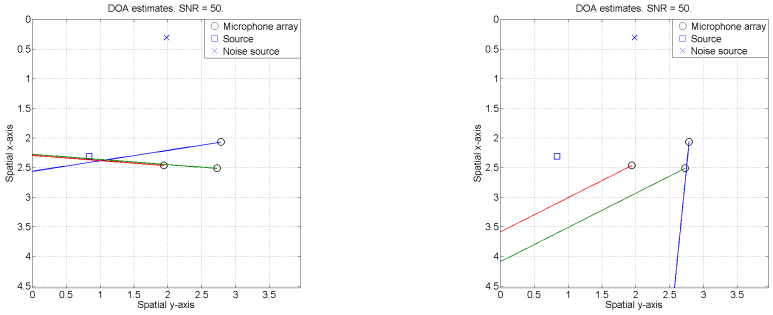
(a) The proposed DOA estimation.



(b) TDOA based DOA estimation [88].

Figure 4.12: DOA estimation in noise field. SNR = 0 dB.

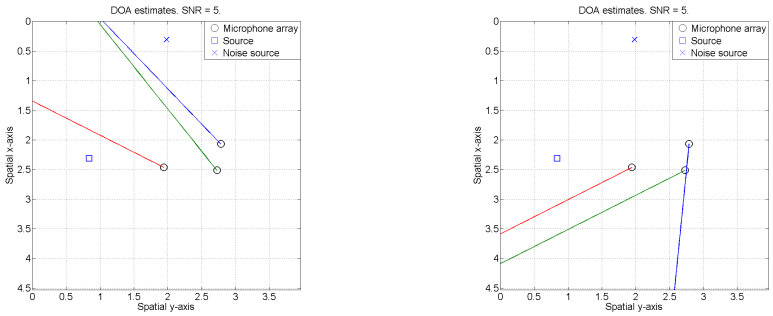
center of the room. The simulation parameters are the same as in the first simulation set, with the exception that the threshold level γ is set to 0.2 dB in both high and low SNR conditions. The DOA was estimated with the proposed estimation method first using the total PBF output power as the beam level ($W = W_s$) estimate and then with PBF SNR as the beam level estimate ($W = W_r$). The results can be seen in Figures 4.13 and 4.14, where SNR of 50 dB and 5 dB are shown.



(a) DOA estimation with the largest output power.

(b) DOA estimation with the largest SNR.

Figure 4.13: DOA estimation. One noise source at 180° with SNR = 50 dB.



(a) DOA estimation with the largest output power.

(b) DOA estimation with the largest SNR.

Figure 4.14: DOA estimation. One noise source at 180° with SNR = 5 dB.

When W_s is used as beam level estimate the DOA estimates do not deviate from the real DOA in high SNR conditions. However, in low SNR the DOA estimates are turned to the direction of the noise source as expected. From this it can be easily concluded that W_s gives good results in high SNR conditions, but the lower SNR more deviated the DOA estimates

are from the real DOA.

As also can be seen from Figures 4.13 and 4.14 when W_r is used, the DOA estimates are turned away from the noise source to provide better SNR. Also, W_r does not differentiate with high and low SNR, but always gives the same DOA estimates if the noise source and the speech source are static. This is because W_r has its maximum when the noise source is at the spatial null of the beam thus attenuating the noise source approximately 15 dB as can be seen from the Figure 4.7(b). This means that in anechoic space (as simulated) the DOA estimation with W_r does not depend on the direction of the speech source but on the minimization of the noise source. In a distributed system where the DOA data of each microphone array is used to localize the speech source this type of behavior is not desired because the error is is even more prominent in the fused data thus significantly reducing the localization accuracy. This is because in the case of W_r the DOA data are dependant of the noise source and not of the speech source as desired.

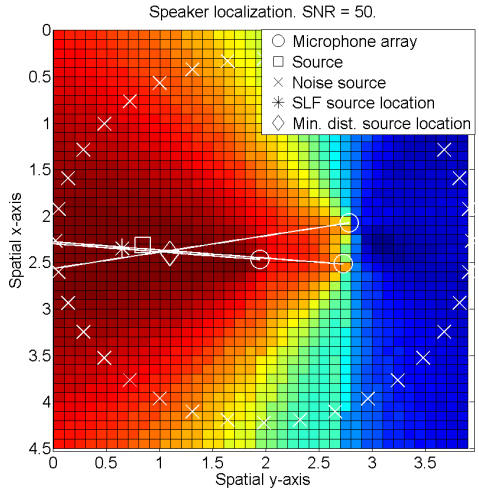
Multiple Arrays

When multiple microphone arrays are used, also the distance of the speech source can be estimated. Here the source locations are estimated with the SLF method described in Chapter 3.3.2 and by searching the point with minimum Euclidean distance from the DOA lines of each array [48]. Similarly as in the previous section, the simulations are done in a noise field by placing uncorrelated Gaussain white noise sources at 10° intervals. The spatial grid of possible source locations is created at 10 cm intervals, which gives total 1840 grid points in the measurement room and random 100 source points are uniformly distributed inside the room. Otherwise the simulation parameters are the same as in the previous section.

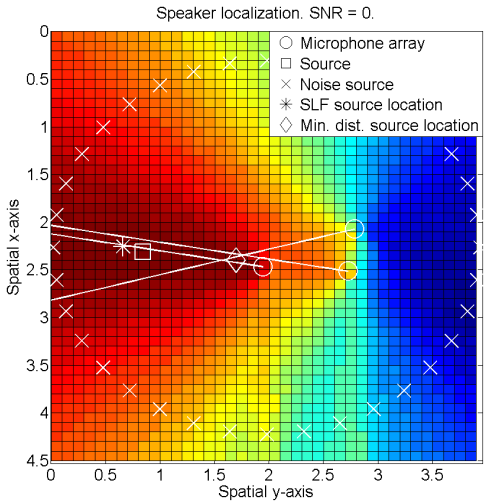
Figure 4.15 shows the speaker SLF of the simulated environment and the estimated source locations when the SNR is 50 and 0 dB at one meter from the source location. The star marks the SLF method localization estimate and the diamond shows the minimum distance source location estimate. The detailed results are presented in Table 4.2, where the mean and the standard deviation of the speaker localization estimates are shown. The results are created by defining 100 random uniformly distributed source locations inside the room and estimating their locations.

Tables 4.2 show that in average both DOA data fusing methods perform similarly in high SNR conditions. However, in the minimum distance method the standard deviation is approximately 85 cm. This is because in the worst-case scenario the DOA lines from closely placed microphone arrays can be almost parallel, resulting in significantly deviated speech source location estimate.

In low SNR conditions the estimates acquired with the SLF based data fusing method



(a) Speaker localization. SNR = 50 dB.



(b) Speaker localization. SNR = 0 dB.

Figure 4.15: Speaker localization in noise field

	SLF		LINE	
SNR	Mean	STD	Mean	STD
50 dB	0.337m	0.193m	0.425m	0.849m
0 dB	0.487m	0.357m	1.396m	0.997m

Table 4.2: The localization error with 100 source points

deviates by approximately 15 cm from the estimates in high SNR conditions. This however is not the case with the minimum distance based method where the average error increases significantly in low SNR conditions. This behavior is expected because although the maximum output power of the each individual PBF would not be steered to the correct speech source DOA, the summing of all the beam powers of all the microphone arrays (3.30) creates a maximum in the global output power of the DMAS.

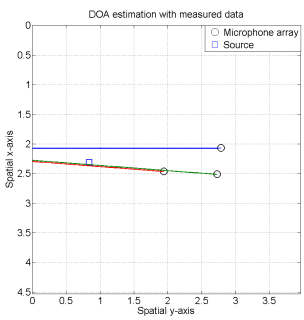
4.4.3 Measured Source Localization Performance

One Array

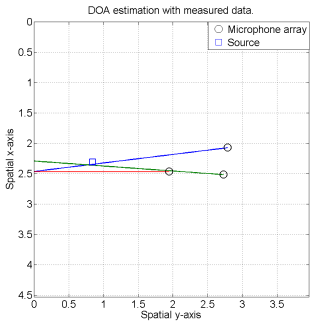
The measurements that were made to evaluate the PBF performance are also used to evaluate the performance of the presented speaker localization framework in a real acoustical environment. Similarly as in the simulated setup, first the DOA of the speech source is estimated with the same two DOA estimation methods. As was noted in Chapter 4.2.5 the measurements were contaminated with a very low frequency noise. Therefore all the frequencies under 300 Hz are ignored when estimating the PBF output power. The cutoff frequency of 300 Hz was chosen because PBF reaches its maximum directivity at 300 Hz as can be seen from Figure 4.7(a).

Figures 4.16 and 4.17 show the DOA estimates of the presented method and the TDOA based method in high SNR conditions and low SNR conditions respectively. The SNR of the microphone output signal was estimated with (4.1) to be approximately 30 dB in high SNR conditions and close to 0 dB in low SNR conditions in 4 kHz bandwidth.

Figures 4.16 and 4.17 show similar performance as in the simulated setup. From Figure 4.17(b) it can be seen that with the measured data the TDOA based method provides better DOA estimations that was expected from the simulations. However, it has to be noted that in simulations the noise signal had the same power at each frequency bin where as in the measured signal the noise has more power in the low frequencies and the high frequencies are not significantly contaminated with noise. Therefore, the cross-correlation based TDOA estimation performs robustly in the high frequencies thus providing good estimates for the TDOA. Also, because most of the noise in the measured data is in the low frequencies and

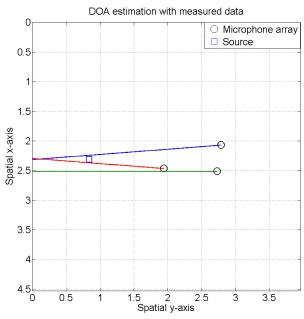


(a) The proposed DOA estimation with measured data

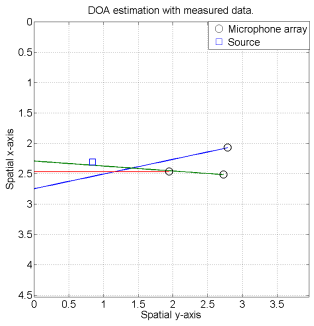


(b) TDOA based DOA estimation [88] with measured data

Figure 4.16: DOA estimation with measured data. SNR $\approx$ 30 dB.



(a) The proposed DOA estimation with measured data

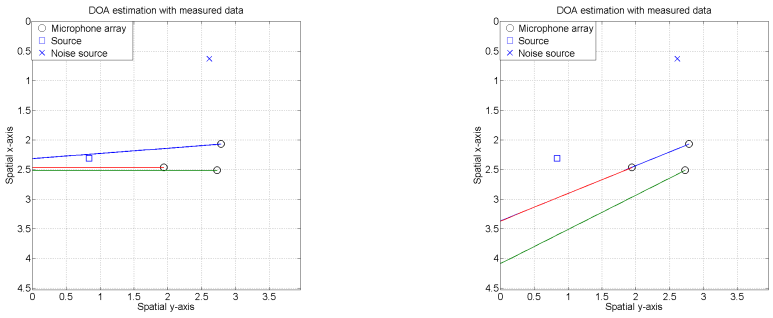


(b) TDOA based DOA estimation [88] with measured data

Figure 4.17: DOA estimation with measured data SNR $\approx$ 0 dB.

for the PBF based DOA estimation method the microphone signals are downsampled to the PBF design sampling rate (8 kHz), the SNR of the signals used in TDOA estimation have approximately 7 dB better SNR than the signals used for the PBF based DOA method.

Figure 4.18 shows the DOA estimation when largest output power is used as the beam level estimate (W_s in Figure 4.18(a)) and when the largest SNR is used as the beam level estimate (W_r in Figure 4.18(b)). Also, white noise is played from the loudspeaker 4 (as defined in Figure 4.3) with sound pressure level of 72 dB measured 30 cm from the loudspeaker. Otherwise, the setting is the same as in the high SNR measurements in the previous chapters.



(a) DOA estimation with the largest output power.

(b) DOA estimation with the largest SNR.

Figure 4.18: DOA estimation. Measured data with white noise from loudspeaker 4.

Figure 4.18 shows similar behaviour as the simulated setup in Figure 4.13. The largest beam power is still at the direction of the speaker and thus also DOA estimates are not significantly deviated from the correct DOA. Also as expected the direction with the highest SNR is not towards the speaker but it is deviated away from the noise source.

Multiple Arrays

As already mentioned when multiple microphone arrays are used, also the speaker location can be estimated. Here the speaker localization framework presented in this work is evaluated with measured data. The processing of the measured data is done by using the same methods and parameters as the simulations in Chapter 4.4.2 in high and low SNR conditions. The speaker localization errors of the presented methods are shown in Table 4.3, where the loudspeaker number refers to the loudspeaker positions as defined in Figure 4.3.

As can be seen from the table, in reverberated conditions the proposed source localization method performs equally well also in low SNR. The table also shows the worst-case-

SNR	**SNR ≈ 30 dB**		**SNR ≈ 0 dB**	
Method	**SLF**	**LINE**	**SLF**	**LINE**
Loudspeaker 1	0.298m	0.080m	0.246m	0.356m
Loudspeaker 2	0.224m	0.874m	0.231m	0.730m
Loudspeaker 3	0.252m	0.840m	0.222m	0.971m
Loudspeaker 4	0.451m	1.825m	0.406m	1.957m
Mean	0.306m	0.905m	0.276m	1.004m
STD	0.101m	0.715m	0.087m	0.684m

Table 4.3: Speaker localization error with measured data

performance of the minimum Euclidian distance method. When the source is at source location 2 although the DOA estimates do not deviate significantly from the real DOA (RMSE < 5°), the DOA lines do not intersect at any point and therefore the source location estimate falls to the boundaries of the room. This is also demonstrated in Figure 4.16(a), where the DOA estimates are plotted when the source is at location 2

When SLF method is used, the small differences in source location estimates can be explained by the averaging of the region with the highest likelihood (equation 3.32). This region changes its location, shape and size depending on the changes in SNR, reverberation, etc. and therefore the source localization can be more accurate in low SNR conditions than in high SNR conditions (Loudspeaker 1, Loudspeaker 3 and Loudspeaker 4). However, more reliable system performance can be achieved by generating a large number of PBF beams thus creating smaller regions with the highest likelihood.

Chapter 5

Conclusions and Future Work

In this work a Distributed Microphone Array System (DMAS) for a two-way audio communication application is presented. The goal of the proposed DMAS is to localize the dominant speech source and capture the speech signal with minimum degradations. Each microphone array in the distributed system works as a Polynomial Beamformer [44] (PBF) and the output of the beamformers is used for the audio capture as well as for the speaker localization. The speaker localization is done by steering the PBF to several possible speech source directions and estimating the signal power or Signal-to-Noise Ratio (SNR) of each beam. The speaker is estimated to be in the steering direction of the beam with highest power or SNR. The fusing of the beam power data of each PBF is done by creating a Spatial Likelihood Function (SLF) by mapping the beam powers of each beam in each microphone array to the acoustical environment and summing these beam powers together. Finally, the speaker is localized to the point with highest value of SLF. The audio capture is done by steering the PBF closest to the speech source to the direction of the localized speaker. This is done in order to maximize the Signal-to-Noise Ratio (SNR) of the PBF output signal.

The proposed two-way audio communication system as well as the DMAS was implemented to run in real-time on top of Pure Data signal processing environment. The implemented communication system uses 32 kHz sampling rate, thus enabling the transfer of nearly full spectrum of human auditory system. To reduce the computational load of the adaptive acoustic echo cancelling algorithm, the adaptive echo cancelling is done only for the low frequencies, while echo suppression techniques are used for higher frequencies. This enables an effective use of adaptive echo cancellation algorithms while at the same time enabling good echo attenuation [80].

The performance of the proposed speaker localization framework was evaluated in simulated anechoic conditions and in real acoustic environment. It was found that the framework was capable of locating the speaker in average error of 40 cm from the real speaker location.

Also, it was shown that the PBF in the implemented microphone array achieves the maximum attenuation of approximately 15 dB for the sounds arriving from the other than the PBF steering direction, thus reducing the effect of reverberation and noise in the captured signal. It was found out that the main restriction in the proposed framework is the finite resolution of the sound source Direction Of Arrival (DOA) estimation. Due to this finite resolution the SLF creates areas of highest values, which change size and shape when even small changes occur in the acoustical environment or in the sound source location. Thus, the proposed method can not be said to perform optimally in any circumstances.

However, the most important theoretical finding in this work is that in a distributed system the use of the beam with the best SNR to determine the speech source direction is not the obvious choice as compared to one array case. In general it can be said that in high SNR conditions using the beams with highest total power provides better results in a distributed system. This is because it was shown that in an environment with one dominant noise source, the Direction Of Arrival (DOA) estimates are dependant of the noise source location and not of the desired speech source location. This leads to larger errors when the deviated data from each microphone array is combined.

Although the results show that the proposed speaker localization framework performs well under static conditions, the performance of the system with moving speakers should be investigated. This might prove to be a challenging task because of the averaging of the power of each beam, which results in slowly reducing power in the beams adjacent to the beam steered to the direction of the active speaker. Also, extensions to multiple active speakers should be investigated.

To enable robust speaker localization in two-way audio communication setup an efficient Voice Activity Detector (VAD) should be implemented to the system. Without a VAD the speaker location might change to the location of the loudspeakers that are used to render the far-end signal or also the speaker location might change even though there is no active speaker. Also, when VAD is placed after the PBF, the VAD could be used to detect if several speakers are active at the same time at different spatial directions.

For the future work in the audio capture side, the formulation of PBF in the frequency domain would enable better estimation of the noise power spectral densities of different spatial directions and those could be used e.g. in spatial noise reduction. Also, Jia *et al.* describe a method combining the beams of each microphone array [41] and thus further increasing the system output SNR, but the benefits of multiple beamformers for audio capture should be more thoroughly investigated.

However, the biggest obstacle in the efficient use of the proposed system in two-way audio communication setup is the constantly changing echo path due to the changes in the steering direction and/or in the location of the microphone array that is used to capture the

signal. When the echo path changes dramatically the echo cancellation algorithm has to use a large step size to adapt itself to the new echo path impulse response which results in audible echo. Especially in multi-speaker conditions the beam steering direction might change rapidly, thus resulting an increasing echo. This problem can be decreased by placing an echo canceller after each microphone array instead of using just one echo canceller. However, this in itself is not enough because not only the active microphone array is changing but also the steering direction of each beamformer. To overcome this problem some kind of echo path history could be used to improve the performance of the echo cancellation. The idea is that every time a beamformer has a new steering direction, it checks if it has been steered there before and retrieves the impulse response of the echo path that was used last time. This behavior would enable the echo canceller to start the adaptation from a close match to the actual echo path impulse response, thus reducing the adaptation time.

However in the end the acceptance of new technologies comes down to the improvements in the user experience of the system. Therefore, listening tests and user experience studies should be performed in order to find out how much beamforming with a distributed microphone array system improves telecommunication experience or if it is enough to have high quality audio hardware components to capture the audio signal.

Bibliography

[1] DPA miniature microphone specifications. http://www.dpamicrophones.com/Images/DM01974.pdf. Last Checked: 16.7.2008.

[2] Genelec 1029a active monitor speaker. http://www.genelec.com/pdf/DS1029a.pdf. Last Chacked: 28.8.2008.

[3] netsend~/netrecieve~ download page. http://www.nullmedium.de/dev/netsend~/. Last Checked: 17.7.2008.

[4] Presonus Firepod user manual. http://www.presonus.com/media/manuals/fpmanual.pdf. Last Checked: 16.7.2008.

[5] Parham Aarabi. The fusion of distributed microphone arrays for sound localization. *EURASIP Journal of Applied Signal Processing (Special Issue on Sensor Networks)*, 2003(4):338 – 347, 2003.

[6] Thibaut Ajdler, Igor Kozintsev, Rainer Lienhart, and Martin Vetterli. Acoustic source localization in distributed sensor networks. In *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, pages 1328 – 1332, November 2004.

[7] Ian F. Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci. A survey on sensor networks. *IEEE Communications Magazine*, 40(8):102 – 114, 2002.

[8] Jason Benesty and Dennis R. Morgan. Frequency-domain adaptive filtering revisited, generalization to the multi-channel case, and application to acoustic echo cancellation. In *Int. Conf. on Acoustics, Speech and Signal Processing*, pages 789 – 792, 2000.

[9] A.J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5):2764 – 2778, 1993.

[10] Stanley T. Birchfield and Rajitha Gangishetty. Acoustic localization by interaural level difference. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1109 – 1112, March 2005.

[11] Joerg Bitzer and K. Uwe Simmer. Superdirective microphone arrays. In Micahel Brandstein and Darren Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 19 – 38. Springer, 2001.

[12] Radamis Botros, Onsy Abdel-Alim, and Peter Damaske. Stereophonic speech teleconferencing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1321 – 1324, April 1986.

[13] Michael Brandstein and Darren Ward. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[14] Herbert Buchner, Robert Aichner, and Walter Kellermann. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Transactions on Speech and Audio Processing*, 13(1):120 – 134, January 2005.

[15] Herbert Buchner, Wolfgang Herbordt, and Walter Kellermann. An efficient combination of multi-channel acoustic echo cancellation with a beamforming microphone array. In *International Workshop on Hand-Free Speech Communication*, pages 55 –58, 2001.

[16] Herbert Buchner and Walter Kellermann. Acoustic echo cancellation for two and more reproduction channels. In *International Workshop on Acoustic Echo and Noise Control*, pages 99 – 102, 2001.

[17] G. Clifford Carter. Tutorial overview of coherence and time delay estimation. In *Coherence and Time Delay Estimation*, pages 1 – 27. IEEE Press, 1993.

[18] Y. T. Chan and K. C. Ho. A simple and efficient estimator for hyperbolic location. *IEEE Transactions on Signal Processing*, 42(8):1905 – 1915, 1994.

[19] Thomas Chou. Frequency-independent beamformer with low response error. In *International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 2995 – 2998, Detroit, USA, 1995.

[20] Dirk Van Compernolle and Stefaan Van Gerven. Beamforming with micophone arrays. In V. Cappellini and A.R. Figueiras-Vidal, editors, *COST 299: Applications of Digital Signal Processing to Telecommunications*, pages 107 – 131. E.U., 1995.

[21] Joseph H. DiBiase, Harvey F. Silverman, and Michael S. Brandstein. Robust localization in reverberant rooms. In Micahel Brandstein and Darren Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 157 – 182. Springer, 2001.

[22] Simon Doclo and Marc Moonen. Gsvd-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing*, 50(9):2230 – 2244, September 2002.

[23] Scott C. Douglas. Blind separation of acoustic signals. In Michael Brandstein and Darren Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 354 –380. Springer, 2001.

[24] Dan E. Dudgeon and Russell M. Mersereau. *Multidimensional signal processing*. Prentice Hall, 1983.

[25] Michael J. Evans, Anthony I. Tew, and James A. S. Angus. Spatial audio teleconferencing - which way is better? In *The 4th International Conference on Auditory Display*, pages 29–38, 1997.

[26] Angelo Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *AES 108th Convention*, Paris, France, February 2000.

[27] C.W. Farrow. A continuosly variable digital delay element. In *IEEE International Symposium on Circuits and Systems*, pages 2641 – 2645, June 1988.

[28] J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, and M.M. Sondhi. Autodirective microphone systems. *Acustica*, 73:58 – 71, 1991.

[29] J.L. Flanagan, J. D. Johnston, R. Zahn, and G.W. Elko. Computer-steered microphone arrays for sound transduction in large rooms. *Journal of the Acoustical Society of America*, 78(5):1508 – 1518, 1985.

[30] Andre Gilloire. Experiments with sub-band acoustic echo cancellers for teleconferencing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 12, pages 2141 – 2144, April 1987.

[31] Lloyd J. Griffiths and Charlse W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27 – 34, 1982.

[32] Timo Haapsaari, Werner de Brujin, and Aki Härmä. Comparison of different sound capture and reproduction techniques in a virtual acoustic window. In *AES 122nd Convention*, Vienna, Austria, May 2007.

[33] Simon Haykin. *Array Signal Processing*. Prentice-Hall, 1984.

[34] Wolfgang Herbordt, Herbert Buchner, Walter Kellermann, Rudolf Rabenstein, Sascha Spors, and H. Teutsch. Full-duplex multichannel communication: real-time implementations in a general framework. In *International Conference on Multimedia and Expo*, pages 49 – 52, 2003.

[35] Wolfgang Herbordt, Herbert Buchner, Satoshi Nakamura, and Walter Kellermann. Multichannel bin-wise robust frequency-domain adaptive filtering and its application to adaptive beamforming. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1340 – 1351, 2007.

[36] Wolfgang Herbordt and Walter Kellermann. Adaptive beamforming for audio signal acquisition. In *Adaptive Signal Processing - Applications to Real-World Problems*, pages 155 – 194. Springer, 2003.

[37] Osamu Hoshuyama and Akihiko Sugiyama. Robust adaptive beamforming. In Michael Brandstein and Darren Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 87 – 109. Springer, 2001.

[38] Yiteng Huang, Jacob Benesty, Gary W. Elko, and Russell M. Mersereau. Rea-time passive source localization: A practical linear-correction leas squares approach. *IEEE Transactions on Speech and Audio Processing*, 9(8):943 – 956, 2001.

[39] Jyri Huopaniemi. *Virtual Acoustics and 3-D Sound in Multimedia Signal Processing*. PhD thesis, Helsinki University of Technology, Department of Electrical and Communications Engineering, 1999.

[40] Matti Hämäläinen and Ville Myllylä. Acoustic echo cancelation for dynamically steered microphone array systems. In *Proc. of the IEEE Workshop of Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2007.

[41] Ying Jia, Yu Luo, Yan Lin, and Igor Kozintsev. Distributed microphone arrays for digital home and office. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 5, May 2006.

[42] Don H. Johnson and Dan E. Dudgeon. *Array Signal Processing*. Prentice Hall, 1993.

[43] Matti Kajala and Matti Hämäläinen. Broadband beamforming optimization for speech enhancement in noisy environments. In *IEEE Worshop on Applications of Signal Processing to Audio and Acoustics*, pages 19 – 22, 1999.

[44] Matti Kajala and Matti Hämäläinen. Filter-and-sum beamformer with adjustable filter characteristics. In *IEEE International Conference on Acoustics and Speech and Signal Processing*, volume 5, pages 2917 – 2920, Salt Lake City, USA, May 2001.

[45] Walter Kellermann. A self-steering digital microphone array. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 3581 – 3584, Toronto, Canada, April 1991.

[46] Walter Kellermann. Acoustic echo cancellation for beamforming microphone arrays. In Michael Brandstein and Darren Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 281 – 306. Springer, 2001.

[47] Zhiyun Li and Ramani Duraiswami. Hemispherical microphone arrays for sound capture and beamforming. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 106 – 109, October 2005.

[48] Hui Liu and Evangelos Milios. Acoustic positioning using multiple microphone arrays. *The Journal of the Acoustical Society of America (JASA)*, 117(5):2772 – 2782, May 2005.

[49] David G. Malham and Anthony Myatt. 3-d sound spatialization using ambisonic techniques. *Computer Music Journal*, 19(4):58 – 70, 1995.

[50] Rainer Martin. An efficient algorithm to estimate the instantaneous snr of speech signals. In *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1093 – 1096, Berlin, Germany, September 1993.

[51] Rainer Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9(5):504 – 512, 2001.

[52] Mar Marzinzik and Birger Kollmeier. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and audio Processing*, 10:109 – 118, 2002.

[53] Robert C. Michael Williamson. Multichannel sound recording practice using microphone arrays. In *24th International Conference: Multichannel Audio, The New Reality*, 2003.

[54] Henrik Møller. Fundamentals of binaural technology. *Applied Acoustics*, 36(3-4):171–218, 1991.

[55] Ville Myllylä. Residual echo filter for enhanced acoustic echo control. *Signal Processing*, 86(6):1193–1205, 2006.

[56] Ville Myllylä and Matti Hämäläinen. Adaptive beamforming methods for dynamically steered microphone array systems. In *IEEE International Conference Acoustics, Speech and Signal Processing*, Las Vegas, USA., Mar. 2008.

[57] Michael Syskind Pedersen, Jan Larsen, Ulrik Kjems, and Lucas C. Parra. A survey of convolutive blind source separation methods. In *Springer Handbook of Speech Processing*. Springer, November 2007.

[58] Jonathan Postel. User datagram protocol. RFC 768, USC/Information Sciences Institute, August 1980.

[59] Jonathan Postel. Transmission control protocol. RFC 793, USC/Information Sciences Institute, September 1981.

[60] Ilyas Potamitis, Huimin Chen, and George Tremoulis. Tracking of multiple moving speakers with multiple microphone arrays. *IEEE Transations on Speech and Audio Processing*, 12(5):520 – 529, September 2004.

[61] Miller Puckette. Pure Data: Another integrated computer music environment. In *Second Intercollege Computer Music Concerts*, pages 37–41, Tachikawa, Japan, 1996.

[62] Miller Puckette. Max at seventeen. *Computer Music Journal*, 26(4):31–43, 2002.

[63] Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society (JAES)*, 45(6):456–466, June 1997.

[64] Ville Pulkki. *Spatial Sound Generation and Perception by Amplitude Panning Techniques*. PhD thesis, Helsinki University of Technology, Department of Electrical and Communications Engineering, 2001.

[65] Boaz Rafaely. Analysis and design of spherical microphone arrays. *IEEE Transactions on Speech and Audio Processing*, 13(1):135 – 143, 2005.

[66] Mika Ristimäki, Matti Hämäläinen, Julia Turku, and Riitta Väänänen. A cross-platform audio signal processing environment for real-time audio algorithm development. In *AES 124th Convention*, Amsterdam, The Netherlands, May 2008.

[67] Gabrielle H. Saunders and James M. Kates. Speech intelligibility enhancement using hearing-aid array processing. *Journal of the Acoustical Society of America*, 102(3):1827 – 1837, September 1997.

[68] H. Sculzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: A transport protocol for real-time applications. RFC 3550, USC/Information Sciences Institute, July 2003.

[69] Xiaohong Sheng and Yu-Hen Hu. Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. *IEEE Transactions on Signal Processing*, 53(1):44 – 53, 2005.

[70] Harvey F. Silverman. Some analysis of microphone arrays for speech data acquisition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(12):1699 – 1712, 1987.

[71] Harvey F. Silverman, William R. Patterson, and James L. Flanagan. The huge microphone array. *IEEE Concurrency*, 6(4):36 – 46, 1998.

[72] K. Uwe Simmer, Joerg Bitzer, and Claude Marro. Post-filtering techniques. In Michael Brandstein and Darren Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 39 – 57. Springer, 2001.

[73] M. Mohan Sondhi, Dennis R. Morgan, and Jospeh L. Hall. Stereophonic acoustic echo cancellation – An overview of the fundamental problem. *IEEE Signal Processing Letters*, 2(8):148 – 151, 1995.

[74] Sascha Spors, Rudolf Rabenstein, and Jens Ahrens. The theory of wave field syntesis revisited. In *AES 124th Convention*, Amsterdam, The Netherlands, May 2008.

[75] Harry L. Van Trees. *Optimium Array Processing*. John Wiley & Sons, 2002.

[76] P. P. Vaidyanathan. Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial. *Proceedings of the IEEE*, 78(1):56 – 93, 1990.

[77] Päivi Valve. Method and system for tracking human speakers. US Patent 6449593, 2001.

[78] Menno van der Wal, Evert W. Start, and Diemer Vries. Design of logarithmically spaced constant-directivity transducer arrays. *Journal of the Audio Engineering Society*, 44(6):497 – 507, 1996.

[79] Barry D. Van Veen and Kevin M. Buckley. Beamforming: Averstile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4 – 24, 1988.

[80] Frederik Wallin and Chrisof Faller. Perceptual quality of hybrid echo canceller/suppressor. In *IEEE International Conference Acoustics, Speech and Signal Processing*, volume 4, pages 157–160, May 2004.

[81] Darren B. Ward, Rodney A. Kennedy, and Robert C. Williamson. Theory and design of braodband sensor arrays with frequency invariant far-field beam patterns. *Journal of the Acoustical Society of America*, 97:1023 – 1034, 1995.

[82] Darren B. Ward, Rodney A. Kennedy, and Robert C. Williamson. Fir filter design for frequency invariant beamfromers. *IEEE Signal Processing Letters*, 3(3):69 – 71, 1996.

[83] Darren B. Ward, Rodney A. Kennedy, and Robert C. Williamson. Constant directivity beamforming. In Michale Brandstein and Darren Ward, editors, *Microphone Arrays:Signal Processing Techniques and Applications*, pages 3 – 18. Springer, 2001.

[84] Darren B. Ward, Eric A. Lehmann, and Robert C. Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on Speech and Audio Processing*, 11(6):826 – 836, 2003.

[85] Eugene Weinstein, Kenneth Sttele, Anant Agarwak, and James Glass. Loud: A 1020-node microphone array and acoustic beamformer. In *International Congress on sound and Vibration (ICSV)*, Cairns, Australia, July 2007.

[86] J.E. West, J. Blauert, and D.J. MacLean. Teleconferencing system using head-related signals. *Applied Acoustics*, 36(3-4):327 – 333, 1991.

[87] Takeshi Yamada, Satoshi Nakamura, and Kiyohiro Shikano. Distant-talking speech recognition based on a 3-d viterbi search using a microphone array. *IEEE Transactions on Speech and Audio Processing*, 10(2):48 – 56, February 2002.

[88] Jari Yli-Hietanen, Kari Kalliojärvi, and Jaakko Astola. Low-complexity angle of arrival estimation of wideband signalsusing small arrays. In *8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing*, pages 109 – 112, 1996.